UNIVERSITY OF ALBERTA

EVALUATING THE EFFECTIVENESS OF TWO-STAGE TESTING FOR ENGLISH AND FRENCH EXAMINEES ON THE SAIP SCIENCE 1996 AND 1999 TESTS

by

©

Gautam Puhan

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements of the requirements for
the degree of Doctor of Philosophy

Department of Educational Psychology
Edmonton, Alberta
Spring, 2003

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisisitons et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 0-612-82159-5
Our file Notre référence
ISBN: 0-612-82159-5

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou aturement reproduits sans son autorisation.

Canada

University of Alberta

Library Release Form

Name of Author: GAUTAM PUHAN

Title of Thesis: EVALUATING THE EFFECTIVENESS OF TWO-STAGE TESTING FOR ENGLISH AND FRENCH EXAMINEES ON THE SAIP SCIENCE 1996 AND 1999 TESTS

Degree: DOCTOR OF PHILOSOPHY

Year Degree Granted: 2003

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

G. Puhan

Gautam Puhan
# 411, 11012-82 Avenue
Edmonton, AB, T6G-2P6

Dec 11, 2002

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, a thesis entitled EVALUATING THE EFFECTIVENESS OF TWO-STAGE TESTING FOR ENGLISH AND FRENCH EXAMINEES ON THE SAIP SCIENCE 1996 AND 1999 TESTS submitted by GAUTAM PUHAN in partial fulfillment for the degree of Doctor of Philosophy.

Dr. Mark J. Gierl

T. Rogers

Dr. W. Todd Rogers

Dr. Mike Carbonaro

Dr. Jacqueline Leighton

Dr. Christina Gagne

T. Rogers

for Dr. Richard Bertrand

Date: Dec. 9/02

DEDICATION

I dedicate this work to my parents Biranchi Narayan Puhan and Arundhati Puhan, my dear sister Bipasha Samal, and my beloved wife Shivani Bhardwaj for their beautiful spirits, love, and support.

Abstract

Two-stage testing (TST) is an adaptive testing procedure where test forms of varying difficulty are administered to examinees based on performance from a routing test. The School Achievement Indicators Program (SAIP) is the national achievement test in Canada and uses the TST procedure to assess educational progress of 13- and 16-year-olds in Science. SAIP works with the implicit assumption that the routing test works equally well for examinees in English and French. If this assumption is true, then there should be proper placement of English- and French-speaking examinees in the second-stage test. However, if the assumption is not true, then there might be misplacement of English- and French-speaking examinees in the second-stage test.

The purpose of the present study was to evaluate the effectiveness of a TST procedure for English- and French-speaking examinees who wrote the SAIP Science tests. The study was conducted using existing data (N=24,642 and 22,320 for the Science 1996 and 1999 administrations, respectively) obtained from SAIP. The analyses were conducted in two steps. First, a comprehensive analysis of the routing test items was conducted using statistical and substantive methods. The purpose of these analyses was to identify items that might favor English- or French-speaking examinees, which in turn, might lead to different placement of examinees in the second-stage test. Second, a comprehensive analysis of the second-stage tests was conducted using BILOG-MG (Zimowski, Muraki, Mislevy, Bock, 1996). To assess performance differences for English- and French-speaking examinees on the second-stage tests, test information functions (TIF), test characteristic curves (TCCs),

standard errors of estimate $SE(\hat{\theta})$, and reliability indices for English and French versions of the second-stage tests were compared.

Statistical analysis of the routing test items revealed that three out of twelve items displayed DIF. However, substantive analysis of the routing test suggested that translation errors were not the cause of DIF for the three items. Analysis of the second-stage tests indicated that English- and French-speaking examinees within low- and high-ability groups performed equally well in the second-stage tests suggesting that the routing test properly placed examinees in the second-stage test for both English- and French-speaking examinees.

Acknowledgements

I would like to thank each of the following people whose help and guidance has made the completion of this dissertation possible.

To Dr. Mark Gierl, my dissertation supervisor, for your interest, knowledge, and insightful suggestions throughout my doctoral program. I appreciated your enthusiasm for good research and your availability and flexibility whenever I needed guidance. I am fortunate and thankful to have such a great mentor.

To Dr. Todd Rogers, for your invaluable comments, suggestions, and directions in carrying out this study. I appreciated your personal attention to detail and for ensuring that the quality of the research met your high standards.

To Dr. Darrell Bock for helping me understand BILOG-MG.

To Dr. Jackie Leighton, Dr. Mike Carbonaro, and Dr. Christina Gagne for your interest and insightful suggestions as members of my dissertation committee.

To Dr. J.P. Das and Dr. David Baine for helping me take some very important academic decisions. Without your help, this endeavor would not have been possible.

To Keith Boughton, a friend and colleague, for your insightful suggestions and ideas throughout my program of research. Your support made everything much easier to accomplish.

Table of Contents

CHAPTER I: INTRODUCTION 1

Purpose of Study 3

CHAPTER II: LITERATURE REVIEW 4

Increased Use of Translated and Adapted Tests 5

Potential Sources of Errors Resulting in Non-Equivalent Language Forms 6

Importance of Test Translations in a Canadian Context 9

Overview of TST 11

Overview of SAIP 15

Context of SAIP 15
Target Groups 16
Sampling 16
Structure 17
Administration 18

TST Framework for SAIP Science 1996 and 1999 Assessment 18

Differential Item Functioning 20

DIF Detection and Interpretation 21

Statistical Methods 21
The SIBTEST Procedure 22
Substantive Methods 25

Differential Item Functioning on Translated Tests 25

IRT-Based Procedures for Monitoring Equivalence 28

Overview of Item Response Theory 29

Item Characteristic Curve 30
Test Characteristic Curve 31

Test Information Function 32
Relative Efficiency 33
Standard Error of Estimate 34
Reliability Index 34

Multiple-Group IRT Theory 35

CHAPTER III: METHOD 37

Overview 37

Method 39

Data 39
Procedure 40
Analysis of the Routing Test 40
Choosing an Appropriate IRT model for the Second-Stage Test Analysis 43
Assessing the Assumptions of Item Response Theory 45
Analysis of the Second-Stage Test 47

CHAPTER IV: RESULTS 52

Analysis of the Routing Test 52

Statistical Analysis 52
Substantive Analysis 53

Analysis of the Second-Stage test 55

Results of the 1996 SAIP Science Test 56
Estimated Latent Distributions: 1996 Administration 56
TIFs for English and French Versions of the Tests: 1996 Administration 57
TCCs for English and French Versions of the Tests: 1996 Administration 57
RE for English and French Versions of the Tests: 1996 Administration 58
SE ($\hat{\theta}$) for English and French versions of the Tests: 1996 Administration 59
Reliability Indices for English and French versions of the Tests: 1996 Administration 59

Results of the 1999 SAIP Science Test 60
Estimated Latent Distributions: 1999 Administration 60
TIFs for English and French Versions of the Tests: 1999 Administration 61
TCCs for English and French Versions of the Tests: 1999 Administration 62
RE for English and French Versions of the Tests: 1999 Administration 63
SE ($\hat{\theta}$) for English and French Versions of the Tests: 1999 Administration 63
Reliability Indices for English and French Versions of the Tests: 1999 Administration 64
CHAPTER V: DISCUSSION AND CONCLUSIONS 67
Summary of Research Questions and Methods 67
Findings 70
Routing Test Analyses 70
Second-Stage Test Analysis 72
Findings from Graphical Procedures 72
Findings from Statistical Procedure 73
Limitations of the Study 75
Recommendations 76
Future Practice 76
Future Research 78
References 81
Appendix A 129
Appendix B 131
Appendix C 132
Appendix D 133
Appendix E 138

List of Tables

Table 1 Results of Item-Guessing Analysis for the Second-Stage Tests: 1996 SAIP Science Administration 88

Table 2 Results of Item-Guessing Analysis for Eight Sub-groups on the Second-Stage Tests: 1999 SAIP Science Administration 89

Table 3 Results of Test-Speededness Analysis for the Easy- and Difficult Second-Stage Tests: 1996 and 1999 SAIP Science Tests 89

Table 4 Summary Statistics for the English and French Versions of the 1996 and 1999 SAIP Science Tests (Routing Test Analysis) 89

Table 5 Results of DIF Analysis using SIBTEST for the SAIP Science Routing Tests: 1996 and 1999 Administrations 89

Table 6 Results of Substantive DIF Analysis for the SAIP Science Routing Test. 89

Table 7 Results of Bundle DIF Analysis using SIBTEST for the SAIP Science Routing Tests: 1996 and 1999 Administrations 89

Table 8 Reliability Indices for English and French Versions of the SAIP Science Second-Stage-Tests for Eight Groups of Examinees: 1996 Administration 89

Table 9 Results of the English and French Comparisons of the TIFs, TCCs, and SE(θ): 1996 SAIP Science Administration 89

Table 10 Reliability Indices for English and French Versions of the SAIP Science Second-Stage-Tests for Eight Groups of Examinees: 1999 Administration 89

Table 11 Results of the English and French Comparisons of the TIFs, TCCs, and SE(θ): 1999 SAIP Science Administration 89

List of Figures

Figure 1 104

Figure 2 105

Figure 3 106

Figure 4 107

Figure 5 108

Figure 6 109

  Panel A 109

  Panel B 109

  Panel C 109

  Panel D 109

Figure 7 110

  Panel A 110

  Panel B 110

  Panel C 110

  Panel D 110

Figure 8 111

  Panel A 111

  Panel B 111

  Panel C 111

  Panel D 111

Figure 9 112

  Panel A 112

  Panel B 112
  Panel C 112
  Panel D 112
Figure 10 113
Figure 11 114
Figure 12 115
Figure 13 116
Figure 14 117
  Panel A 117
  Panel B 117
  Panel C 117
  Panel D 117
Figure 15 118
  Panel A 118
  Panel B 118
  Panel C 118
  Panel D 118
Figure 16 119
Figure 17 120
Figure 18 121
Figure 19 122
Figure 20 123
Figure 21 124

Figure 22 125
  Panel A 125
  Panel B 125
  Panel C 125
  Panel D 125
Figure 23 126
  Panel A 126
  Panel B 126
  Panel C 126
  Panel D 126
Figure 24 127
Figure 25 128

CHAPTER I: INTRODUCTION

Assessment is an important method for monitoring student achievement. Assessment programs are increasingly used to evaluate student performance on well-defined, problem-solving and curricular-related tasks. For this reason, it is important that assessments are fair for all examinees (*Principles for Fair Student Assessment Practices for Education in Canada*, 1993).

In the context of fairness, the effectiveness of translated tests, as tools that yield scores that can be validly interpreted regardless of the language group to which they belong, is often questioned (*Standards for Educational and Psychological Testing*, 1999). A poor translation can affect the meaning of test items and adversely influence the comparability and interpretability of test scores across language and cultural groups (e.g., Hambleton & Patsula, 1999; Ercikan, Gierl, McCreith, Puhan, & Koh, 2002). For instance, a reading test developed in English that is translated to French may include content not equally meaningful or appropriate for students who read only French. Gierl and Khaliq (2001, p.175) provide an example in which an item with a contour relief map contained the phrase "cross section cut along a line" in the English form while in the French form contained the phrase " une coupe transversale qui montre le relief." The idea of relief is excluded from the English form. This difference could seriously affect the comparability of the two forms and disrupt the intended purpose of the test (e.g., comparison and interpretation of test scores across the two language groups).

Achievement tests are often translated and adapted for use in different languages and cultures. Therefore, the translation process must be accurate. For example, the Council of Ministers of Education in Canada assesses the achievement of 13- and 16-year-old students in reading and writing, mathematics, and science in *English and French* as part of the Council's School Achievement Indicators Program (SAIP). Similarly, the Department of Learning in the province of Alberta translates eight of their 11 English high school exit exams into French. Hambleton (1993) and Sireci (1996) highlight the need to enhance the fairness of comparisons for individuals and groups from different language and ethnic backgrounds. These comparisons must be fair because there is a growing interest in cross-cultural research resulting in comparative studies across national, ethnic, and cultural groups. (for recent examples, see Hambleton & Patsula, 1998; Jeanrie & Bertrand, 1999; Reckase & Kunce, 1999).

The issues that surround the proper methodology for adapting a test to support valid and reliable comparisons of scores are complex. The effects of translation errors may even be compounded when complex testing procedures, like two-stage testing (TST), are used to assess student performance on achievement tests. TST is a testing procedure in which test forms of varying difficulty are administered to examinees based on previous performance estimated from a first-stage test, commonly referred to as the *routing* test (Zimowski, Muraki, Mislevy, & Bock, 1996). The routing test is one of the most important features of the TST procedure. If the routing test has items with translation errors, then it may not route examinees from different language groups equally well. Consequently, there can be misplacement of examinees in the *second-stage test*, which is unfair.

SAIP uses a of a two-stage testing (TST) procedure. The results are used to compare the performance of English- and French-speaking examinees in the subject areas assessed. The implicit assumption made is that the routing test works equally well for both English- and French-speaking examinees. If this assumption is true, then English- and French-speaking examinees with the same ability will be placed at a similar location on the score scale at the second-stage test. If this assumption is not true, then there might be misplacement of English- and French-speaking examinees in the second-stage test. For example, if the routing test has items that are biased and favor French-speaking examinees, then more French-speaking examinees will be routed to a high-ability test even though they should have been routed to a low-ability test. Consequently, the French-speaking examinees will take a more difficult second-stage test and may perform poorly compared to the English-speaking examinees. Such an outcome may be considered unfair because it could adversely affect the reported achievement levels of French-speaking examinees. There is a well-documented body of research regarding the problems of translating and adapting tests to different languages (e.g., Hambleton & Patsula, 1998; Van deVijver & Leung, 1998). However, the adverse effect of translated tests on assessment instruments used in a two-stage testing design is not well documented.

Purpose of Study

The Council of Ministers of Education, Canada uses a TST procedure to compare the performance of 13-and 16-year-old, English- and French-speaking examinees. Since fairness is an important concern in the field of educational measurement, it is necessary to ensure that a particular form of testing, such as TST, does not unfairly favor one language group compared to another. Hence, the purpose of the present study is to compare the performance of English and French examinees that wrote the SAIP Science 1996 and 1999 achievement tests administered using the TST procedure. Four research questions are addressed:

1. Is there evidence for differential item performance for English- and French-speaking examinees on the routing test used in the TST procedure?
2. If so, what is the source of differential item performance in the routing test?
3. Is there evidence for differential test performance for English- and French-speaking examinees on the second-stage test used in the TST procedure?
4. Is there a relationship between performance on the routing test and the second-stage test used in the TST procedure?

The evaluation of whether assessment procedures such as TST produce comparable results for different groups of examinees such as English and French requires consideration of key psychometric concepts such as test translation and adaptation, equivalence, and differential item functioning (DIF). These concepts are reviewed in the next chapter.

CHAPTER II: LITERATURE REVIEW

Increased Use of Translated and Adapted Tests

Achievement tests are often adapted for use in different languages and cultures. For example, the International Association for the Evaluation of Educational Achievement (IEA) conducted the Third International Mathematics and Science Study in 1995 and 1996. Altogether, the tests were administered in 31 different languages to students in 45 participating countries. Similarly, the Organization for Economic Co-operation and Development (OCED) conducted the Programme for International Student Assessment (PISA) in 2000. A test of reading literacy, mathematical literacy, and scientific literacy were administered in different languages to students in 32 participating countries. In Canada, the Council of Ministers of Education assesses achievement of 13- and 16-year-old students through the School Achievement Indicators Project (SAIP) in reading and writing, mathematics, and science in English and French. Other prominent examples of test adaptations include Spanish versions of the College Board's Scholastic Assessment Test (SAT), the American Council on Education's General Educational Development Tests (GED), and the United States Department of Education's National Assessment of Educational Progress (NAEP)(see Hambleton & Patsula, 1998, p.1).

Hambleton and Patsula (1998) and Sireci (1996) list a number of reasons to explain the increased interest in test adaptations and translations. These reasons include an increase in international exchanges of tests, more demand for credentialing and licensure exams in multiple languages, a reduction in cost because of adapting an existing test as compared to constructing a new test, the need to develop and translate tests to certify employees in their native language, and a growing interest in cross-cultural research. In the Canadian context an increased interest in test translations and adaptations may be attributed to increased number of international tests, the introduction and continuation of the national tests administered by SAIP, and the testing programs of several provinces in which tests are administered in both official languages, English and French.

Potential Sources of Errors Resulting in Non-Equivalent Language Forms

The process of translating and adapting tests into different languages and cultures is often viewed as an easy task. Typically the translation process only requires finding someone who knows the languages and that no more than a couple of hours are needed to translate the test (Hambleton & Patsula, 1999). However, translating a test from a source language to a target language does not necessarily produce two psychometrically equivalent tests (Allalouf et al. 1999). Angoff and Cook (1988, p. 2) wrote: "It can hardly be expected without careful and detailed checks, that the translated items will have the same meaning and relative difficulty for the second group as they had for the original group before translation."

Many problems are cited in the psychometric literature to explain why translated tests may not be equivalent with the original test. Hambleton (1993, 1994) identified six problems, which are described below.

First, the construct measured in the source language may change when the test is translated into a second language. According to Hambleton (1994), it is entirely possible that the same construct is interpreted and understood in completely different ways in two different languages or cultural groups. For example, the Western notion of intelligence places considerable emphasis on speed of response. However, in many Eastern cultures, speed of response is unimportant and intelligence is often associated with thoughtfulness, reflection, and saying the right thing (Lonner, 1990).

Second, there are many concepts, expressions, and ideas used in the source language version that do not have equivalents in the target language. For example, Gierl and Khaliq (2001) noted that the English version of an item included the sentence, "Most rollerbladers favor a helmet bylaw." The word rollerblader has no equivalent word in the French language, thus making it difficult to translate the test item so that it conveyed the same meaning in English and French.

Third, the meaning of an item can change during test translation. Hambleton (1994, p. 235) provides one illustrative example. In a Swedish-English comparison, English-speaking examinees were presented with this item:

Where is a bird with webbed feet most likely to live?

a. in the mountains

b. in the woods

c. in the sea

d. in the desert.

In the Swedish translation the phrase "webbed feet" became "swimming feet" thereby providing a visible clue to the Swedish-speaking examinees about the correct option for this item. This example provides an instance of *construct-irrelevant easiness* by providing extraneous cues (e.g., link between sea and swimming) in the item that permitted the Swedish population to respond correctly in ways irrelevant to the construct being measured (Messick, 1989, p. 35).

Fourth, greater cultural distance between language and ethnic groups adversely affects test equivalence. For example, tests may show more comparability across similar language groups such as French- and Spanish-speaking groups as compared to English- and French-speaking groups (W. Todd Rogers, June 2002, personal communication).

Fifth, comparability of translated tests is often a function of what is measured. For example, tests that require knowledge obtained from school may show less comparability across different cultural and language groups than a test of memory span (Van de Vijver & Leung, 1997).

Sixth, cultural differences regarding test administration procedures often pose a threat to test equivalence across different language and cultural groups. According to Hambleton (1994), administration procedures range from language use in test rubrics, lay-out and use of graphics, presentation mode (e.g., paper and pencil, computer), and specific formats (e.g., multiple choice, essay). One frequent problem identified by Hambleton and Patsula (1999) centres on the presentation format of the test, which may be less familiar to persons in one culture than another. For example, the authors note that the multiple-choice format is very familiar in North America but less common in other parts of the world. Also in many Eastern cultures, due to limited access to computers, the paper-and-pencil mode of taking tests is more common than computer-based testing.

These six problems illustrate how differences between languages and cultures may adversely affect the validity of translated and adapted tests. Hambleton (1994) notes that since "high-stakes" are often associated with the results from cross-cultural comparisons or international comparative studies of educational achievement, the need for *test equivalence* across different language and cultural groups becomes extremely important. Moreover, the adverse consequences of bias in translated assessment materials are immense (Hambleton, 2001). Improper translations may make the test instruments easier for students in some cultural or language groups but not in others. Similarly, extended and/or simplified translations may make the test instruments easier for students in other sub-cultures and languages.

Importance of Test Translations in a Canadian Context

Concerns with the accuracy of test translations and adaptations should be particularly important for test developers and users in bilingual countries like Canada since many national and provincial tests must be administered in the official languages of the country. Many of these tests are first developed in English and then translated to French. Hence, it becomes necessary to ensure that both the English and French versions of these tests are comparable and that no particular language form (e.g., English) unfairly favors a particular group (e.g, English-speaking examinees).

Canada's national testing program, the School Achievement Indicators Program (SAIP), is used to assess 13-and 16-year-old students in the areas of reading, writing, mathematics, and science in English and French. The examinations in this testing program are developed in both English and French and are designed to be equivalent for the English- and French-speaking examinees. However, little information is presented in the SAIP reports about the bilingual test development process beyond the fact that both English- and French-speaking test developers were involved in writing the test items with the intent of eliminating any linguistic bias

(Council of Ministers of Education, Canada, 2000, p. 8). Unfortunately, little research has been presented to support the assumption that the English and French forms used by SAIP are parallel despite the importance of this topic for valid test score interpretation and use.

The effects of translation errors may even be compounded when complex testing procedures, like two-stage testing (TST), are used to assess student performance on achievement tests. This topic is particularly important in the context of SAIP because this testing program uses a TST procedure to scale test scores and compare performance of English- and French-speaking examinees in various subject areas. For example, in the SAIP Science assessments, a 12-item first-stage test or the routing test is used to route examinees to an appropriate second-stage test form based upon the ability level of the examinees derived from the routing test. SAIP scores their tests and reports results with the implicit assumption that the routing test works equally well for both English- and French-speaking examinees. If this assumption is true, then English- and French-speaking examinees with the same ability will be assigned to the same test at the second stage. If this assumption is not true, then there might be misplacement of English- and French-speaking examinees in the second-stage test. For example, if the routing test has items that are biased and favor French-speaking examinees, then more French-speaking examinees will be routed to the test form for a high-ability group even though they should have been routed to the test form for a low-ability group. Consequently, the French-speaking examinees will take a more difficult second-stage test and may perform poorly compared to the English-speaking examinees. Such an outcome may be considered unfair because it could adversely affect the reported achievement levels of French-speaking examinees. Hence, the purpose of the present study was to evaluate the effectiveness of a TST procedure for English- and French-speaking examinees who wrote the SAIP Science 1996 and 1999 tests.

Test development for SAIP Science was comparable for the 1996 and 1999 administrations, hence replication of findings across years can help cross validate the results. The comparability of SAIP Science 1996 and 1999 administrations is evident from the following description: *Changes to assessment instruments and scoring procedures across the 1996 and 1999 administrations were kept to a minimum. The same Framework and Criteria was used to assess student work. Scoring procedures and conditions as well as administration procedures for the 1999 Science administration were replicated as much as possible from documentation and information provided by the Science 1996 team* (Council of Ministers of Education, Canada, 2000, p. 6).

In order to describe the context and methods relevant to this study, literature will be reviewed in the following four areas: two-stage testing (TST), School Achievement Indicators Program (SAIP), differential item functioning (DIF), item response theory or IRT (including IRT for multiple groups), and IRT-based methods for monitoring test equivalence.

Overview of TST

TST is a form of adaptive testing suitable for group administration. It works by tailoring the difficulties of the test forms to the abilities of selected groups of examinees (Zimowski, et al. 1996). TST begins with a preliminary estimate of examinees' ability obtained from a short first-stage test or the routing test. Based on the results of the routing test, examinees are classified into different levels of abilities (e.g., low, moderate, and high). Depending on the examinee's estimated ability derived from the routing test, examinees are then assigned to a second-stage test form with a difficulty level consistent with their estimated ability level. The basic idea underlying TST is that lower-ability examinees should perform relatively poorly on the routing test and can therefore be directed to an easier second-stage test. High-ability examinees, in contrast, should score well on the routing test and should be directed to a more difficult second-stage test.

An important consideration in TST is that the scoring of a two-stage test must take into account the systematic differences in difficulties of the second-stage tests. Hulin, Drasgow, and Parsons (1983, p.212) give an example of a hypothetical two-stage test where examinee A answered 10 items correctly on the most difficult second-stage test and examinee B answered 10 items correctly on the easiest second-stage test. According to the authors, the number right score of examinees A should not be compared to the number right score of examinee B because examinee A answered 10 difficult items correctly (and incorrectly answered the remaining difficult items) and examinee B answered 10 easy items correctly (and incorrectly answered the remaining easy items). Hence, to allow a direct comparison of the number right score of examinee A and the number right score of examinee B, the second-stage test forms are linked with a set of common items so that they can be calibrated on the same score scale extending from the lowest to the highest levels of ability (for a more detailed discussion of linking procedures, see Zimowski et al, 1996; Sireci, 1996; Angoff & Cook, 1988).

The key element in TST is the routing test because it is performance on this test that determines which of the second-stage tests an examinee will be assigned. Examinees can be assigned to an inappropriate second-stage test if the routing test does not place different examinees equally well for the language groups to be compared. Routing error can result from either routing too high or too low. Routing too high would occur when an examinee is assigned to a second-stage test that is above the examinee's ability. Similarly, routing too low would occur when an examinee is assigned to a second-stage test that is lower than the examinee's ability. In the context of group comparisons in testing, if examinees in a particular language group (e.g., English or French) are misrouted, then examinees of that group may be able to pass the second-stage test when they should have failed or may fail the second-stage test when they should have passed.

To be effective, a routing test should be short and consist of a set of representative and unbiased items that are highly discriminating and have a spread of difficulty values for the different language groups (Bock & Zimowski, 1998). However when there is a large number of content areas, it may be difficult to have a short routing test. Developing items that are not equally difficult and discriminating can also be a difficult task.

Despite these difficulties, TST has great potential for application in achievement testing. Bock and Zimowski (1998) list some benefits of TST in achievement testing. First, TST as compared to conventional paper-and-pencil testing results in better measurement precision, especially in the tails of the proficiency distribution in the population of examinees because the second-stage tests are tailored to the ability levels of the examinees, hence reducing measurement error.

Second, TST can be cost effective as compared to conventional paper-and-pencil testing because of the reduction in the number of items required to estimate IRT scale scores with equal precision.

Third, administering tests in CAT format is very expensive because of the need for a sufficient, often large, number of computers for administering these tests. Furthermore, development of an effective and robustly operating item bank to obtain reliable estimates of examinees' proficiency scores is costly. Paper-and-pencil TST could provide for a more cost-effective alternative as compared to CAT. Also paper-and-pencil TST could play an important role in transition to a fully developed computerized system.

In large-scale assessments where results are considered low-stakes for the examinees, use of TST can be beneficial due to a number of factors. First, lack of student motivation is a major problem in low-stakes examinations such as SAIP. TST can result in better motivation for students during the testing session because it avoids presenting discouragingly difficult items or very easy items.

Second, TST has the potential to increase the usability and validity of results such as obtained from SAIP by making available good-quality student level scores in different subject matter areas (see Bock & Zimowski, 1998, p.24).

Third, TST enables the conditioning on provisional estimates of student proficiency within the adaptive session. Bock and Zimowski (1998) note that conditioning on provisional estimates is much stronger than conditioning on background characteristics.

These benefits indicate potential of TST in present and future assessments at national and provincial levels. However, as in any new implementation of a measurement instrument for assessing performance on well-defined problem solving and achievement tasks, it is important to ensure that the interpretations drawn from the measurement instrument are reliable and valid. In the *Principles for Fair Student Assessment Practices for Education in Canada* (1993), guideline seven states "Assessment instruments translated into a second language or transferred from another context or location should be accompanied by evidence that inferences based on these instruments are valid for the intended purpose (p. 7)." For example, if the routing test functions differentially for different language groups such as English and French, then the inferences based on the second-stage tests may be invalid for comparing performance of English- and French-speaking examinees.

Overview of SAIP

Context of SAIP

SAIP is a national testing program in Canada designed to monitor educational progress of 13- and 16-year-old students. It is conducted by the Council of Ministers of Education (CMEC). It is comparable to the National Assessment of Educational Progress (NAEP) project which reports on U.S. student performance with comprehensive information about what students at grades four, eight, and twelve know and can do in various subject areas.

SAIP serves four major purposes or uses. First, students are assessed on tests designed to measure "what students in Canadian schools are expected to know and be able to do" (CMEC, 2000, p. 6). What is to be assessed is determined by evaluation and curriculum specialists from universities, content experts, and representatives from non-governmental organizations. Second, the test results are used to determine whether students in each province and territory in Canada achieved similar levels of performance in literacy, numeracy, and science at about the same age. Third, the results are used to determine whether students in different provinces and territories in Canada learn similar skills for solving problems in reading and writing, mathematics, and science. Fourth, it complements existing assessments in each province and territory by providing Canada-wide data on the achievement levels attained by students across the country.

Target Groups

The SAIP assessments are administered to 13- and 16-years old students. The group of 13-year-olds consists of students in their first year of secondary school, which is the transition year between elementary and secondary school. The group of 16-year-olds consists mainly of students in their last year of compulsory school attendance.

Sampling

The samples of 13- and 16-year-olds are selected using a two-stage procedure. At the first stage, schools with 13- and 16-year-old students are selected from the list of all schools provided by the territories and provinces in Canada to the CMEC. Schools under federal jurisdiction and those with fewer than five students are excluded. Then, the schools are selected using a procedure that takes their size into account. At the second stage, students are selected using a uniform procedure that requires the coordinators for the provinces and territories to allow equal probability of selection of all 13- and 16-year-olds.

Structure

The SAIP tests consist of a written assessment and a hands-on performance assessment. The written assessment includes both multiple-choice and written-response questions that measure the acquisition of concepts, procedures, and problem-solving skills. Both multiple-choice and written-response questions in the written assessment are scored dichotomously, one for a correct response and zero for a wrong response. In the hands-on performance assessment students are required to demonstrate levels of performance related to their inquiry skills. Questionnaires are also given to students, teachers, and principals to gain information on demographic and psychological variables, attitudes towards a particular subject, opportunities to learn a particular subject, and type of instruction. The data from the written- and hands-on performance assessments and from the background questionnaires are used to provide descriptions of students' strengths and weaknesses in basic and higher-order skills; compare achievement by race/ethnicity, gender, type of community, and region; describe trends in performance across years given these tests are administered approximately every three years, and determine the relationships between achievement and background variables (e.g., homework, employment, reading materials in the home, TV watching) and instruction (e.g., amount of instructional time and hands-on-learning).

Administration

The first cycle of assessments began in 1993 with the administration of the mathematics assessment in April of that year. The reading and writing assessment followed in 1994. The first science assessment was administered in 1996. A random-sample of students is assessed in these content areas once, every three years with one exception, in which, there was a gap of two years between the first and second Science assessments.

TST Framework for SAIP Science 1996 and 1999 Assessment

Students writing the SAIP Science assessments in 1996 and 1999 were administered a first-stage test consisting of 12 items. Based on the results of the first-stage test, examinees were classified into one of two levels of ability. The low-ability group consisted of examinees who scored seven or lower in the first-stage test. The high-ability group consisted of examinees who scored eight or above in the first-stage test. The first-stage test was followed by assignment of the low-ability group to an easy second-stage test and the high-ability group to a more difficult second-stage test. Each second-stage test consisted of 66 questions that covered a different combination of achievement levels ranging from one (lowest) to five (highest). The easy second-stage test contained 26 level one questions, 26 level two questions, and 14 level three questions. The difficult second-stage test was composed of 14 level three questions, 26 level four questions, and 26 level five questions. The 14 level 3 questions were identified in both forms so as to allow the placement of the scores obtained on the two forms on a common scale. The description of different achievement levels along with example items are presented in Appendix A.

In a TST framework, the first-stage test corresponds to the routing test. As emphasized earlier, the routing test is the key element in TST. If the routing test consists of items that are not highly discriminating and are not free from the effects of differential item functioning (DIF) then there can be misplacement of examinees in the second-stage testing and the effectiveness of the TST procedure can be compromised. SAIP works with the implicit assumption that the routing test works equally well for examinees in different language groups such as English and French. This assumption is based upon the fact that from the outset, the instruments used in the science assessment were developed by English- as well as by French-speaking educators working together for the purpose of eliminating any possible linguistic bias. If this assumption is true then English- and French-speaking examinees will be placed in a similar manner to the second-stage test which, in turn, will lead to English- and French-speaking examinees performing equally well in the second-stage test. However, if this assumption is not true then there might be misplacement of English- and French-speaking examinees in the second-stage test (e.g., high-ability French examinees taking an easy test and low-ability English examinees taking a difficult test) which might be unfair.

The evaluation of whether SAIP Science assessments are comparable for English- and French-speaking examinees requires a consideration of statistical and substantive procedures used to study the comparability of scores across different language and cultural groups. A brief overview of differential item functioning or DIF is presented in the next section followed by a discussion of DIF as applied to test translation.

Differential Item Functioning

DIF is present when examinees from different groups have a different probability of answering an item correctly after conditioning on overall ability (Shepard, Camilli, & Averill, 1981). The total test score is most frequently used as an estimate of ability. DIF methods match examinees on total test scores to see if comparable examinees from different populations (e.g., English-speaking versus French-speaking examinees who have the same total test score) perform the same on individual items. If they do not perform the same, then the item is said to display DIF. However, DIF does not mean simply that an item is harder for one group than for another; if the students in one group tend to know more than the other group about the subject, they will tend to perform better on the items in a test. This is referred to as item impact (Clauser & Mazor, 1998).

DIF analysis is one of the most important statistical analyses in validating a test score for use in different cultural and language populations in an item bias study (Hambleton, 1994). Support for comparability or equivalence of a test for two groups comes from the fact that when members of two groups have equal ability, then they should perform in an equivalent manner on each item. The main idea, as discussed earlier, is that if members of two groups have equal ability, then their performance on each item in a test should be equal except for sampling errors due to sample size. However, if group differences in performance beyond sampling errors are noted, the item is labeled DIF and more intensive investigations are carried out to identify the source of these differences (see Clauser & Mazor, 1998 for a review).

DIF Detection and Interpretation

Statistical Methods

A variety of statistical procedures for detecting DIF have been developed (Camilli & Shepard, 1994; Millsap & Everson, 1993; Fidalgo, 1996a; Potenza & Dorans, 1995). Of these, the Mantel-Haenszel (MH), Simultaneous Item Bias Test (SIBTEST), and Logistic Regression (LR) have been the most commonly used (e.g., Allalouf et al. 1999; Gierl & Khaliq, 2001; Gierl, Rogers, & Klinger, 1999). Moreover, of these procedures, SIBTEST has been found to be more effective than MH and LR in detecting DIF (Jiang & Stout, 1998; Bolt & Stout, 1996, Gierl et al. 1999). SIBTEST has two well-documented benefits. First, SIBTEST uses a regression estimate of the true score instead of the observed score to match students with the same ability. As a result, examinees are matched on a latent rather than an observed score. Second, SIBTEST can be used to assess DIF iteratively by initially using all the items from the matching test and systematically removing DIF items from the matching test until a subtest of items without DIF is identified (Shealy & Stout, 1993). Furthermore Gierl et al. (1999) and Ercikan et al. (2002) have shown that SIBTEST identifies more DIF items as compared to MH and LR. From a test-development point of view, detection of more DIF items may be problematic because of the enormous costs incurred to develop test items. However, from an interpretation point of view, identification of more DIF items may result in a more thorough analysis of the test items leading to a more comprehensive test interpretation. Therefore, SIBTEST was used in the present study to identify items with DIF. In the next section an overview of SIBTEST is presented.

The SIBTEST Procedure

SIBTEST is a non-parametric method, which was developed as an extension of Shealy and Stout's (1993) multidimensional model for DIF. In the SIBTEST framework, DIF is conceptualized as a difference between the probabilities of selecting a correct response, for examinees with the same levels of the latent attribute of interest (θ). This difference, when found, is attributable to different amounts of nuisance abilities (η) that influence the item response patterns.

The statistical hypothesis tested by SIBTEST is:

$$H_0\colon B(T) = P_R(T) - P_F(T) = 0$$

versus

$$H_1\colon B(T) = P_R(T) - P_F(T) \neq 0,$$

where $B(T)$ is the difference in probability of a correct response on the studied item for examinees in the reference (or advantaged) and focal (or disadvantaged) groups matched on true score; $P_R(T)$ is the probability of a correct response on the studied item for examinees in the reference group with true score T; and $P_F(T)$ is the probability of a correct response on the studied item for examinees in the focal group with true score T. With the SIBTEST procedure, items on the test are divided into two subsets, the suspect subtest and the matching subtest. The suspect subtest contains items that are suspected of having DIF and the matching subtest contains items that, ideally, are known to be unbiased and measure only the primary dimension on the test. Linear regression is used to estimate the corresponding subtest true score for each matching subtest score. These estimated true scores are adjusted using a regression correction technique to ensure the estimated true score is comparable for the examinees in the reference and focal groups on the matching subtest (Shealy & Stout, 1993). In the final step, $B(T)$ is estimated using, $\hat{B}_{UNI}$, which is the weighted sum of the differences between the proportion-correct true scores on the studied item for examinees in the two groups across all score levels. The weighted mean difference between the reference and focal groups on the studied subtest item or bundle across the *k* subgroups in given by

$$\hat{\beta}_{\text{UNI}} = \sum_{k=0}^{k} p_k d_k \,,$$

where p_k is the proportion of focal group examinees in subgroup *k* and d_k is the difference in the adjusted means on the studied subtest item or bundles of items for the reference and focal groups, respectively, in each subgroup *k*.

SIBTEST provides an overall statistical test and a measure of the effect size ($\hat{B}_{UNI}$) for each item. $\hat{B}_{UNI}$ has a standard normal distribution with mean 0 and standard deviation 1 under the null hypothesis of no DIF. A statistically significant value of $\hat{B}_{UNI}$ that is positive indicates DIF against the focal group and a negative value indicates DIF against the reference group. There is one major advantage of having both a statistical test as well as a measure of effect size ($\hat{B}_{UNI}$) for identifying DIF items. An effect size measure can be of great importance where Type I error inflation can pose serious threats to the validity of results derived from a statistical test. For SIBTEST, when Type I error occurs due to statistically biased DIF estimation, it is important to estimate whether the biased estimation is small or large.

For instance, a highly inflated Type I error might not be a serious problem when the $\hat{B}_{UNI}$ is close to zero (Roussos & Stout, 1996).

Research at the Educational Testing Service (ETS) has resulted in guidelines for classifying DIF as negligible, moderate, and large using the Mantel-Hanzel statistical procedure. Roussos and Stout (1996, p. 218, p. 220) adopted the ETS guidelines and applied the results to SIBTEST. According to the authors the following $\hat{B}_{UNI}$ values are used for classifying DIF:

No DIF: Null hypothesis is not rejected and $|\hat{B}_{UNI}| \cong 0$,

Negligible or Level A DIF: Null hypothesis is rejected and $|\hat{B}_{UNI}| < 0.059$,

Moderate or Level B DIF: Null hypothesis is rejected and $0.059 \leq |\hat{B}_{UNI}| < 0.088$,

Large or Level C DIF: Null hypothesis is rejected and $|\hat{B}_{UNI}| \square 0.088$.

A comprehensive and technical discussion of the SIBTEST procedure is found in Shealy and Stout (1993).

In the context of the routing test used in the SAIP Science assessments, statistical methods such as SIBTEST can be used to identify items that function differentially for English- and French-speaking examinees. As discussed earlier, if the routing test has items that function differentially for English- and French-speaking examinees, then it might lead to misplacement of examinees in the second-stage test. Hence, identification of such items is necessary to improve the effectiveness of the TST procedure.

Substantive Methods

Statistical methods, as described earlier, are only useful in detecting DIF items. To understand the nature of DIF, judgmental reviews are often used to identify *why* items are functioning differentially between groups (see, for example, Camilli and Shepard, 1994, p. xiii; Gierl & Khaliq, 2001). According to Hambleton (1994), items flagged as DIF may be problematic because of poor translation or because of the use of a term or expression that is unknown or unfamiliar to the examinees in one of the groups. The skill measured by the translated item may not be part of the repertoire of the target language population. Alternatively, the difference in performance may be due to systematic inherent differences between the examinees from the two groups. Determining the reason for the difference is important because it influences the ultimate decision of what to do with the item. Without substantive analysis, it will be difficult to know whether an item that is flagged statistically as DIF is the result of translation errors or actual differences between the examinees from the two populations (e.g., bias versus impact).

Differential Item Functioning on Translated Tests

Researchers who study the psychometric characteristics of translated tests have noted the presence of large amounts of DIF items on many translated tests. For example, Gierl et al. (1999) reported that 26 of 50 items (52%) on a Canadian Grade 6 Social Studies achievement test translated from English to French displayed moderate or large DIF. Similarly, Allalouf et al. (1999) noted that 42 out of 125 verbal items (34%) displayed moderate or large DIF on the Israeli Psychometric Entrance Test when Hebrew and Russian examinees were compared. A more recent study conducted on SAIP tests by Ercikan et al. (2002) revealed that for English- and French-speaking examinees, approximately 18 to 31 percent of the items in reading, 32 to 37 percent of the items in mathematics, and 32 to 36 percent of the items in Science showed DIF for 13- and 16-year-olds. These findings raise questions about the validity of translated tests for assessing achievement outcomes. According to Gierl and Khaliq (2001), such findings also highlight the need to identify potential sources of translation DIF so that necessary steps can be taken during test development to improve the items.

Gierl and Khaliq (2001) note that if specific sources of translation errors could be anticipated, then test developers could carefully monitor their test construction, translation, and adaptation practices to ensure that different language forms of the exam are comparable across language groups. The authors outline four major sources of translation errors that might affect validity of tests across different language groups.

First, omission or additions of words, phrases, or expressions that affect meaning are likely to affect performance for one group of examinees. For example, on an item with a contour relief map, the English form contained the phrase "cross section cut along a line" while the French version contained the phrase "une coupe transversale qui montre le relief". The idea of relief is included in the French version but not in the English version of the test and this difference might adversely affect the performance of a particular language group.

Second, differences in words, expressions, and structure of sentences of items that are inherent to the language and/or culture might affect the performance of a particular group of examinees. One example that illustrates a cultural difference includes an English item with a 12-hour clock using a.m. and p.m. while the French version uses a 24-hour clock. This time convention is common between the English and the French. Students were required to interpret a time difference when solving this item. Gierl and Khaliq (2001) found this item to consistently favour French-speaking examinees. Without taking translation into consideration, one might conclude that the French-speaking examinees have a better understanding of time differences as compared to English-speaking examinees. However, when translation differences are accounted for, it appears that the French-speaking examinees have an advantage on this item because the 24-hour clock makes the correct option more salient.

Third, differences in words, expressions, and sentence structure of items that are not inherent to a language or culture might affect the performance of a particular group of examinees. For example, Gierl and Khaliq (2001, p.175) identified an item on an Alberta achievement test that contained the phrase in English "traditional way of life" versus the phrase in French "les traditions." This item presented two distinct concepts surrounding "a way of life" and "traditions" in the English and French forms, respectively and this difference might adversely affect the performance of a particular language group.

Fourth, differences in item and test format such as change in item or test structure, typeface, capitalization, and punctuation may affect the performance of one group of examinees. If these differences provide a clue to the correct answer for one group of examinees, then the item is not comparable across language groups (for a more detailed list of the sources of translation problems, see Gierl & Khaliq, 2001; also see Allalouf et al. 1999).

These four specific sources of translation errors show how differences in words, phrases, ideas, punctuation, and item structure in different languages and cultures can adversely affect comparability of test scores for different language groups. Allalouf et al. (1999) note that more research should be conducted to identify specific sources of translation errors because if these sources could be predicted, then they could be taken into account at an early stage in the test development process, thus resulting in improved decisions regarding test construction, scoring, and equating (e.g., excluding items that function differentially across languages in an equating design).

IRT-Based Procedures for Monitoring Equivalence

Measures based on item response theory (IRT) are considered among the best methods for evaluating item and test equivalence (e.g., Drasgow & Hulin, 1991; Thissen, Steinberg, & Gerrard, 1986). Commonly used IRT procedures for determining equivalence of tests for different groups of examinees include comparison of item characteristic curves (ICC-equivalence), test characteristic curves (TCC- equivalence), test information functions (TIF- equivalence), relative efficiency (RE- equivalence), and standard errors of estimate (SE- equivalence) (Boughton, 2001). A brief overview of IRT is presented in the next section followed with a discussion of commonly used IRT-based procedures for determining test equivalence and multiple-group IRT theory. It should be noted that these IRT procedures are conceptualized within the two-parameter (2PL) IRT model because the 2-PL model is used in the present study. The rationale for use of the 2-PL model is presented in Chapter III.

Overview of Item Response Theory

It is postulated in IRT that for any examinee and test item interaction there is an underlying ability or proficiency level that influences performance on that item. Examinee performance is therefore a function of both item and person characteristics. An examinee with a high level of ability has a greater probability of answering an item correctly than an examinee with a low level of ability. This relationship between ability and item performance is usually described by a monotonically increasing nonlinear function called item characteristic curve (ICC). These curves provide the probability of getting an item correct at each ability level across the entire ability or theta scale. Thus, with an accurate estimate of ability, subsequent examinee performance can be predicted (Hambleton & Swaminathan, 1985).

There are three common assumptions underlying the use of unidimensional IRT models: unidimensionality, local independence, and speededness of response. The first assumption, unidimensionality, assumes that there is only one underlying trait or ability that accounts for an examinee's response to a test (e.g., science ability). The second assumption, local independence, states that an examinee's responses to different items must be statistically independent. That is, the order of the items administered must not affect person's performance on the test. If the assumption of unidimensionality is met, then it automatically results in local independence among the items. The third assumption, speededness of response, requires that all examinees have enough time to attempt all items so that only their ability level affects their responses to each item and not a failure to reach an item. It is implicit in the unidimensional model that only one dimension is being measured and not a second ability (e.g., speed) in answering a question (Hambleton & Swaminathan, 1985).

Within IRT, there are three popular models that provide a mathematical equation for the relation of probability of correct response to ability. The first model, the one-parameter logistic model (or Rasch model), assumes that all items have equal discriminating power and that guessing is zero. The single item parameter that is estimated is the difficulty or *b*-parameter. The second model, the two-parameter (2PL), is a more general model than the one-parameter logistic model and assumes that items vary in both difficulty (b_i) and discrimination (a_i). The third model, the three-parameter logistic model (3PL) was introduced in order to account for a lower ability examinee obtaining the correct response by chance to a difficult item. Hence, the 3PL model assumes that items vary in difficulty (b_i), discrimination (a_i), and a third parameter know as the pseudo guessing parameter, c_i (for a more detailed discussion of IRT, see Hambleton, Swaminathan, & Rogers, 1991).

Item Characteristic Curve

ICCs provide a means for comparing the responses of two different groups (e.g., English versus French) to the same item. According to Lord (1980), two ICCs can be different only if the item parameters that describe them are different. A difference in ICCs of two groups indicate that examinees from an English- or French-speaking group at the same ability level do not have the same probability of success on the item which, in turn, might result in DIF for the item. Under the two-parameter logistic model, the item characteristic curve (ICC) is calculated using the equation given by Birnbaum (1968)

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}.$$

where D = 1.7, $P_i(\theta)$ = the probability that a randomly chosen examinee with ability θ answers item *i* correctly, b_i = the item *i* difficulty parameter, and a_i = the item *i* discrimination index. In Figure 1, the ICCs for the English and French versions of a hypothetical item are compared. As seen in Figure 1, the item is more difficult for French-speaking examinees than English-speaking examinees. With an increase in ability there is a greater probability of correct response for the English speaking examinees than for the French-speaking examinees.

Test Characteristic Curve

Test characteristic curves (TCCs) provide a means for comparing the responses of two different groups (e.g., English versus French) to the same test. The TCC, which is the sum of the ICCs at a given ability level, is calculated using the equation

$$T = \sum_{i-1}^{n} P_i(\theta).$$

A difference in TCCs for two groups indicates that examinees from an English- or French-speaking group at the same ability level do not have the same total test score which, in turn, might indicate that the source test and the target test are not equivalent. In Figure 2, TCCs for the English and French version of a hypothetical test are compared. As seen in Figure 2, the test is more difficult for French-speaking examinees than English-speaking examinees because with an increase in ability, the total test score for the English-speaking examinees is greater than the French-speaking examinees.

Test Information Function

Test information functions (TIFs) provide a method for comparing the responses of two different groups (e.g., English versus French) to the same test. A difference in TIFs of two groups indicate that examinees from an English- or French-speaking group at the same ability level do not provide the same amount of information about the test which, in turn, might indicate that the source test and the target test are not equivalent. Under a two-parameter model, the item information function (IIF) for items in each group is calculated using the equation (see Lord, 1980)

$$I_i(\theta) = D^2 a_i^2 P_i Q_i,$$

where D = 1.7, a_i = is the item discriminator parameter, P_i = the probability of an examinee at a certain theta (θ) level obtaining the correct answer to item i, and $Q_i = 1 - P_i$.

The TIF, which is the sum of the IIFs at a given ability level, is calculated using the equation

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta).$$

In Figure 3, TIFs of the English and French version of a hypothetical test are compared. As seen in Figure 3, the test provides more information for English-speaking examinees than French-speaking examinees at all points on the ability scale.

Relative Efficiency

A comparison of the information function from two tests will also give an indication of whether or not the tests are approximately equivalent. Relative efficiency (RE) makes a comparison of the information function of one test (e.g., English test) with a second test (e.g., French test) in terms of a common ability scale. In all RE comparisons a straight line is interpreted as a theoretical baseline, a line where the RE is one or the same for examinees from two language groups such as English and French. As the RE is the ratio of information for the English-speaking over the French-speaking examinees, a value of RE greater than one means more information for English-speaking examinees and a value less than one means more information for the French-speaking examinees. The formula for RE is

$$RE(\theta) = \frac{I_E(\theta)}{I_F(\theta)},$$

where RE(θ) denotes the relative efficiency and $\frac{I_E(\theta)}{I_F(\theta)}$ denotes the information of the English test compared to the French test over a common ability θ. In Figure 4, the relative efficiency of test A is compared with test B (assuming both tests have 30 items each). In this example, test B is functioning as if it were 20% shorter than test A at a theta of around 0.5. This outcome means that we would need to increase test B from 30 to 36 items in order to produce precision of the ability estimates of test A. Thus, relative efficiency will aid in the assessment of whether tests are equivalent by allowing a direct comparison of source language tests with the target language tests (Hambleton et al. 1991).

Standard Error of Estimate

The standard error of estimate, which is the reciprocal of the square root of the test information function, can provide information on the degree of precision at which an ability level is estimated. For example, if at a particular ability level the standard error is higher for English-speaking examinees than French-speaking examinees then it can be said the ability level is more precisely estimated for French-speaking than English-speaking examinees. This outcome, in turn, might indicate that the English test and the French test are not equivalent. $SE\ (\hat{\theta})$is calculated using the equation

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}},$$

where $I(\theta)$ is the test information function. In Figure 5, the standard error of estimate is compared for English and French versions of a hypothetical test. As seen in Figure 5, the standard error is lower for French-speaking examinees than English-speaking examinees at the lower end of the ability scale. This outcome indicates that at the lower end of the ability scale, the ability level of French-speaking examinees is estimated more precisely than the English-speaking examinees.

Reliability Index

The reliability index indicates consistency with which a test measures a particular ability (e.g., science ability). Because similar reliability indices may indicate similar performance for different language groups, the reliability index can be used to assess whether English- and French-speaking examinees are performing equally well on a particular test (e.g., science test). The reliability indices for different groups of examinees is calculated by using the equation

$$R = 1 - ME,$$

where R = reliability index and ME = measurement error of variance. According to Bock and Zimowski (1998), ME = $\frac{1}{TIF}$. Thus, for example, a value of 5 for TIF corresponds to a ME of $\frac{1}{5}$ = 0.2 and a reliability index of 1- 0.2 = 0.8.

Multiple-Group IRT Theory

To ensure that scores are comparable across groups when different groups of examinees complete different sets of items, analyses based on multiple-group IRT are often conducted in which scores for multiple groups of examinees are scaled to a common metric using common items or examinees (Bock & Zimowski, 1998). Multiple-group analyses are particularly useful when examinees are grouped on the basis of one or more variables (e.g., language and ability) and then compared. For example, language groups are often sub-divided according to ability levels and age to include more homogenous groups of examinees in the analysis.

In multiple-group IRT, it is assumed that the different groups of examinees are drawn from populations that have score distributions that are normal but have different means and standard deviations. Under these assumptions the item response data can be estimated completely by estimating the means and standard deviations of the groups along with the item parameters. In a two-stage testing situation, where the groups correspond to examinees who have been selected on the basis of a first-stage or routing test, the latent distributions of the second-stage groups cannot be considered normal even when the latent distributions of the populations from which the examinees were selected are normal. In such cases, the empirical distributions can be estimated along with item parameters by maximum marginal likelihood. The indeterminacy of location and scale is resolved, either by setting the means and standard deviations of one of the groups to convenient values, such as zero and one, or by setting the overall mean and standard deviation of the combined distribution to similar values (for an example of multiple-group IRT application, see Bock & Zimowski, 1998).

To summarize, tests that are fair must provide comparable information for diverse groups of examinees such as English and French. This outcome is often questionable on translated tests when examinees from different language groups are compared because poor translation can affect the meaning and interpretability of particular constructs between translated tests (e.g., Gierl, in-press). Consequently, the validity of translated tests for comparing achievement outcomes across different language groups (e.g., English- and French-speaking examinees) becomes an important psychometric topic. The purpose of this study is to compare the performance of English- and French-speaking examinees who wrote the SAIP Science 1996 and 1999 achievement tests administered using a two-stage testing procedure.

CHAPTER III: METHOD

Overview

Evaluating the effectiveness of a TST procedure for English- and French-speaking examinees who wrote the SAIP Science 1996 and 1999 tests requires an analysis of the routing test and the second-stage test. If the assumption that the English and French versions of the routing test tests are equivalent is true, then it is reasonable to expect English- and French-speaking examinees to perform equally well on the second-stage test. However, if the routing test does not place English- and French-speaking examinees equally well, then English- and French-speaking examinees of differing ability can be assigned to the same second-stage test, possibly leading to performance differences for these two language groups in the second-stage test. The evaluation of the effectiveness of TST for English- and French-speaking examinees, therefore, requires a comprehensive analysis of the routing test and the second-stage test.

Empirical DIF methods can be used to detect items in the routing test that function differentially for English- and French-speaking examinees. If the routing test has items that are free from the effects of DIF for English- and French- speaking examinees, then it will route English- and French-speaking examinees equally well to the second-stage test. Likewise, comparable performance of English- and French-speaking examinees on the second-stage test would indicate that the routing test placed examinees from these two language groups equally well to the second-stage test.

The routing-test analysis was conducted separately for 13- and 16-year olds for the 1996 and 1999 SAIP Science achievement tests. The second-stage analyses were also conducted separately for 13- and 16-year olds within the low- and high-ability groups for the 1996 and 1999 SAIP Science achievement tests. The replication of analyses for two age and ability groups across two populations contributes to the evaluation of the effectiveness of two-stage testing for English- and French-speaking examinees in three ways: (a) the cross-validation of findings from one age group with another, (b) the cross-validation of findings from one ability group with another and, (c) the cross-validation of findings from one population with another.

The current study was conducted in two steps. In the first step, a comprehensive analysis of the routing test items was conducted using statistical and substantive methods. The purpose of these analyses was to identify items that might favor English- or French-speaking examinees which, in turn, might lead to different placement of examinees in the second-stage test. Statistical analyses for the routing test included identifying DIF items using SIBTEST. Substantive analyses included using test translators who helped identify potential sources of translation errors.

In the second step, the performance of English- and French-speaking examinees for two age and ability levels on the second-stage tests was compared using the following procedures:

1. Test information functions (TIFs) were compared for English and French versions of the test,
2. Tests characteristic curves (TCCs) were compared for English and French versions of the test,

3. Relative efficiency of the French version compared with the English version of the test was computed,
4. Standard errors of estimate, denoted as $SE(\hat{\theta})$ were compared for English and French versions of the test, and
5. Reliability indices were calculated for the English and French versions of the test to obtain as estimate of the effectiveness of the TST procedure in general.

The two step approach was used in the present study because results from the routing test analysis and the second-stage test analysis can provide information about the effectiveness of the two stage testing procedure. If the routing test has items that function differentially for English- and French-speaking examinees, then it can lead to misplacement of examinees in the second-stage test. Comparison of the English and French versions in the second-stage analysis either supports or rejects the above assumption.

Method

Data

The study was conducted using existing data collected in the 1996 and 1999 administrations of Science. Eight distinct groups of examinees, classified by ability (high and low), age (13- and 16 yrs), and language (English and French), were analyzed. In the 1996 Science administration, the low ability group was composed of 13-year-old English (n=5,171) and French (n=1,986) examinees and 16-year-old English (n=2,772) and French (n=1,101) examinees. The high ability group was composed of 13-year-old English (n=4,347) and French (n=1,540) examinees and 16-year-old English (n=5,713) and French (n=2,012) examinees. In the 1999 Science administration, the low ability group was composed of 13-year-old English (n=4,431) and French (n=1,549) examinees and 16-year-old English (n=2,178) and French (n=858) examinees. The high ability group was composed of 13-year-old English (n=4,086) and French (n=1,449) examinees and 16-year-old English (n=5,729) and French (n=2,040) examinees.

There are two main reasons for having eight distinct subgroups in the analysis. First, there is a forced dichotomy in the data because examinees were categorized as a high ability or low ability group based on their performance on the first-stage test. Hence, the high ability group took only the difficult second-stage test and the low ability group took only the easy second-stage test. Second, there might be actual differences in the construct being measured (also know as impact) between 13- and 16-year-old examinees. This, in turn, would make it difficult to interpret whether performance differences between English- and French-speaking examinees were a result of translation DIF or impact or both. Hence, the groups were divided into 13- and 16-year-olds so that the analyses could be conducted using more homogenous subgroups.

Procedure

Analysis of the Routing Test

A comprehensive analysis of the routing test was conducted using statistical and substantive methods. Statistical analyses for the routing test included using SIBTEST to identify items that function differentially for English- and French-speaking examinees. DIF analyses were conducted on each item from the English and French forms of the routing test. For each suspect item, the remaining items were used as the matching subtest. Recall from Chapter II that the matching subtest should, ideally, consist of items that are unbiased and free from the effects of DIF. Since the routing test has only 12 items, the matching subtest may become seriously contaminated if there are many DIF items in the routing test. Hence, for more valid interpretations of the results of the DIF analysis, the 14 common level 3 items in booklet B and booklet C in the second-stage tests were also included in the DIF analysis of the routing test. Since these 14 items were taken by all examinees, inclusion of these items in the DIF analysis increased the number of items in the matching subtest from 12 to 26, leading to a more stable set of results.

SIBTEST provides an overall statistical test and a measure of the effect size for each item ($\hat{B}_{UNI}$ is an estimate of the amount of DIF). According to Roussos and Stout (1996, p. 220) the following $\hat{B}_{UNI}$ values are used for classifying DIF:

- No DIF: Null hypothesis is not rejected, and $|\hat{B}_{UNI}| \cong 0$,
- Negligible or Level A DIF: Null hypothesis is rejected and $|\hat{B}_{UNI}| < 0.059$,
- Moderate or Level B DIF: Null hypothesis is rejected and $0.059 \leq |\hat{B}_{UNI}| < 0.088$,
- Large or Level C DIF: Null hypothesis is rejected and $|\hat{B}_{UNI}| \square 0.088$.

These guidelines were used to classify DIF items in the present study.

In all English-French comparisons, items with a B- or C-level rating were considered DIF items whereas those with an A-level rating were not considered as DIF items. This decision seems justified since B- and C-level DIF items are typically scrutinized for potential bias in tests reviews (Zieky, 1993). Also, in all analyses, an alpha level of 0.05 was used with a non-directional hypothesis test.

For the substantive analyses, test translators and item writers identified potential translation errors. A translation review process developed by Gierl and Khaliq (2001) was used in the present study. Four bilingual French-English translators completed a blind review of the routing test items. They were asked to identify items with translation problems. The four translators were native French speakers. They were also completely bilingual in the English and French languages. The four translators had extensive experience in teaching ranging from seven to twenty-three years. The translation review process not only required the identification of differences in the two language versions but judgments regarding whether the differences were expected to lead to performance differences for the two language groups as well. Therefore, experience in teaching and familiarity with student thinking processes were important skills for the translators.

In the translation review process, the four translators first worked separately. They were asked to evaluate the similarities and differences between the English and French test items in the routing test. For each item, the four translators were asked to specify which language group would be favored, identify the exact reason or reasons for the difference, if present, in each item, and then categorize the reason or reasons for the difference into the four sources of the translation errors identified by Gierl and Khaliq (2001) (see p. 26). The four translators were also asked to create their own sources of translation error if they found the sources identified by Gierl and Khaliq (2001) to be insufficient. Once the task was completed, the four translators met to discuss their decisions. The meeting allowed each translator to defend his or her decision for every item and the test translators as a group, to reach consensus on the items where they disagreed. The routing test items were reviewed in one meeting. The review process required two hours and forty-five minutes in order to reach consensus across the four translators.

Choosing an Appropriate IRT model for the Second-Stage Test Analysis

As in all IRT applications, a necessary first step is to choose an appropriate item response model. An initial classical test theory (CTT) analysis of the items in the second-stage tests had shown that the discrimination indices varied considerably for the items. For the 1996 SAIP Science data, the range for the discrimination indices was 0.57 for the easy second-stage test and 0.56 for the difficult second-stage test. For the 1999 Science data the range for the discrimination indices was 0.65 for the easy second-stage test and 0.48 for the difficult second-stage test. With this prior information about the discriminating powers of the items there was no reason to fix the value of $a = 1$ for all items. Hence, the one-parameter logistic IRT model was not chosen. Furthermore, of the two- and three-parameter logistic IRT models, item parameters could not be estimated using the three-parameter logistic IRT model because of non-convergence of the conditional marginal maximum likelihood estimates. Hence, the three-parameter logistic IRT model was not used. Consequently, the two-parameter logistic IRT model was chosen for both conceptual and empirical reasons. First, almost 40 % of the items used in the second-stage tests were constructed-response items and guessing is considered to have minimal influence on examinees performance for constructed-response items. Therefore, using a three-parameter model for all the items was not considered appropriate. Second, the second-stage tests in TST are tailored to the abilities of the examinees. Therefore it was reasonable to assume that the guessing parameter will be considerably low for the multiple-choice items. Hence, using a two-parameter model for these items was considered appropriate. Third, Bock and Zimowski (1998) suggested that for very large sample sizes such as in the present study (N=24,642 and 22,320 for the Science 1996 and 1999 administrations, respectively), small differences in model fit would be detected by the use of different IRT models. Fourth, the performance of low-scoring examinees (based on their total score) was examined on the most difficult items. The expectation was that the low-scoring examinees would have close to zero performance on the most difficult items if the assumption of no guessing is true. For both the 1996 and 1999 administrations of the SAIP Science assessment, examinees who scored less than one third of the total score in the second-stage tests were considered as low scoring examinees (Ndalichako & Rogers, 1997). Their performance on the most difficult items was examined. Analysis was conducted separately for the eight sub-groups used in the present study. The results of the item-guessing analyses are presented in Tables 1 and 2 for the 1996 and 1999 SAIP Science tests, respectively. As seen in Tables 1 and 2, for the eight sub-groups, the performance of the low scoring examinees on the three most difficult items was close to zero suggesting minimal guessing in the second-stage tests. For example, as seen in Table 1, for the low-ability, 13-year-old English group, only two out of 689 examinees who scored less than one third on the total test got item 35 correct.

Also, from an empirical point of view, the two-parameter logistic IRT model was used because of convergence of the conditional marginal maximum likelihood estimates. Based on the above reasons, the two-parameter model was used throughout the present study.

Assessing the Assumptions of Item Response Theory

Model selection can be aided by an investigation of the principal assumptions underlying the popular unidimensional item response models. Two important assumptions common to all these models (discussed earlier in Chapter III) are that the data are unidimensional and the test administrations are not speeded.

A linear factor analysis was conducted to assess the dimensional structure of the SAIP Science tests. Many researchers have stressed that real test data often cannot be well modeled using strictly unidimensional models (e.g., Ackerman, 1987, 1989; Ansley & Forsyth, 1985; Yen, 1985). This is particularly true for achievement test data, which often have a dominant first factor and a cluster of items that indicate the existence of other dimensions. The problem of achieving strict unidimensionality with real data have led researchers to assess unidimensionality based on the concept of essential unidimensionality, which attempts to model the presence of a dominant dimension in the presence of minor dimensions (Stout, 1990, Nandakumar, 1991). Hence, in the present study, *unidimensionality* of the second-stage tests was assessed using the concept of essential unidimensionality.

A principal components analysis of tetrachoric correlations was conducted. The analysis was conducted separately for the eight sub-groups used in the present study. The eigenvalues (from largest to smallest) of the inter-item correlation matrix were studied to determine whether a dominant first factor for the second-stage tests was present. For the 1996 Science test, the scree plots for the easy- and difficult second-stage tests for the eight sub-groups are shown in Figures 6 and 7, respectively. Similarly, for the 1999 Science test, the scree plots for the easy- and difficult second-stage tests for the eight sub-groups are shown in Figures 8 and 9, respectively. As seen in Figures 6 through 9, the scree plots for the second-stage tests for the eight sub-groups studied suggest the presence of a dominant first dimension which, clearly dominates the remaining components.

For the 1996 SAIP Science administration, the eigenvalues for the easy- and difficult second-stage tests for the eight sub-groups studied are presented in Appendix B. Similarly, for the 1999 Science administration, the eigenvalues for the easy- and difficult second-stage tests for the eight sub-groups studied are presented in Appendix C. As seen in Appendix A and B, the eigenvalue of the first component in both the easy and difficult second-stage tests is considerably larger than the eigenvalue of the remaining components. For example, as seen in Appendix B, for the low-ability, 13-year-old English group, the eigenvalue of the first component is 9.65, which is more than three times larger than the eigenvalue of the second component. Further, the difference in eigenvalues for the first and second components is considerably larger than the difference in eigenvalues for the second and third components. This result held true for both 1996 and 1999 SAIP Science tests for the eight sub-groups studied.

Results of the linear factor analysis show that while not large in an absolute sense, the magnitude of the first component in both second-stage tests is large enough to indicate that there is a dominant component underlying the item responses (Huynh

& Ferrara, 1994). This, consequently suggests that the assumption of essential unidimensionality is met.

To assess *speededness of response*, the percentages of examinees completing 75% of the test were reviewed to determine whether or not the second-stage tests were speeded (Hambleton et. al, 1991). Results of the analysis to test speededness of the second-stage tests for the eight sub-groups studied are presented in Tables 3 for the 1996 and 1999 SAIP Science tests. As seen in Table 3, for both the 1996 and 1999 second-stage tests, a considerably high percentage of examinees completed 75 % of the items in the easy- and difficult second-stage tests within the eight sub-groups studied. For example, as seen in Table 3, for the 1996 SAIP Science test, 93.40 % of 13-year-old English examinees within the low-ability group completed 75 % of the items in the easy second-stage test. Hence, speed is assumed to be an unimportant factor.

Analysis of the Second-Stage Test

All computations in the analysis were conducted using the BILOG-MG program (Zimowski, Muraki, Mislevy & Bock, 1996). The program command files for the analysis of the 1996 and 1999 data are included in Appendix D and E, respectively.

The analysis was carried out in two stages. The first stage consisted of estimating the routing test item parameters and latent distributions. The second stage consisted of estimating the link and second-stage item parameters and the latent distributions. For the second-stage analysis, the latent distributions estimated in the routing test or first-stage analysis were used as the prior distributions for maximum marginal likelihood estimation of the combined routing test and second-stage test data. The latent distributions show the proficiencies of different groups of examinees in a particular subject matter (e.g., knowledge about science). These distributions are usually reported in terms of the means and standard deviations (SD) of the achievement levels for different groups of examinees. Although the case records from the different forms (e.g., English and French forms) are subjected to a single IRT item analysis in BILOG-MG, the test form is identified on each case and separate latent distributions are estimated for examinees taking different forms. The points and weights representing the distributions for the 1996 and 1999 data are shown in the BILOG-MG command files in Appendices D and E, respectively.

IRT scaling of items in the second-stage tests was done using the 14 common items. These items correspond to the level 3 items described earlier (p. 18). These items served as the link items between the second-stage tests and have average difficulty and high discrimination indices. The reason for scaling the items in the second-stage tests using IRT is to put the item parameters of items in the two second-stage tests on a common metric (Angoff & Cook, 1988; also see section on multiple-group IRT in chapter II). This, in turn, allows comparison of the item parameters in the English and French versions of the second-stage tests for different groups of examinees (e.g., 13-year-old, English- versus French-speaking examinees).

To assess performance differences for English-speaking and French-speaking examinees on the second-stage tests, the test information functions (TIF) for English- and French- versions of the second-stage tests for different groups of examinees were first compared. Second, test characteristic curves (TCC) for English and French versions of the second-stage tests for different groups of examinees were compared. Third, the information functions of the English- and French- versions of the second-stage tests were compared by computing the relative efficiency of the French and English tests. Fourth, standard errors of estimate, denoted as $SE(\hat{\theta})$ were compared for English- and French- versions of the test. Fifth, reliability indices were calculated for the English- and French- versions of the test to obtain as estimate of the effectiveness of the TST procedure in general. Similar reliability indices indicate equivalence across the second-stage test forms.

The use of more than one IRT based procedure for monitoring test equivalence is justified because different IRT methods give slightly different information about the psychometric properties of the test. For example, TCCs and TIFs can be used to describe properties of a test. Both curves illustrate identical data, but the value of the TCC is more sensitive to variations in the *b*-parameter of the test items while the TIF is more sensitive to variations in the *a*- and in case of the 3-PL logistic IRT model, the *c*- parameters of the items. Hence, by evaluating the value of the TCC and the TIF, we will have a more accurate understanding of how much any tests differ.

The TIFs, TCCs, and the $SE(\hat{\theta})$ for the second-stage tests were compared using two procedures. First, a graphical procedure was used in which the TIFs, TCCs, and SEs obtained for different groups of examinees who took the second-stage tests (e.g., low ability French-versus English-speaking examinees) were compared visually and an estimate of the magnitude of difference between two empirical curves was obtained. However, the graphical method provided only a rough estimate of the magnitude of difference between two empirical curves. Therefore, a statistical procedure was used to obtain more precise estimates of the magnitude of difference between two empirical curves. The mean square residual (MSR) was calculated for each pair of TIFs, TCCs and SEs (e.g., high ability French-versus-English-speaking examinees). The equation used to calculate MSR is

$$MSR = \frac{\sum_{1}^{n}\left[X_i(\theta) - Y_i(\theta)\right]^2}{n-1},$$

where n represents the number of score points on the theta scale and in the context of comparison of TIFs, $X_i(\theta)$ = value of information at θ for test 1 (e.g., English version of the test) and $Y_i(\theta)$ = value of information at θ for test 2 (e.g., French version of the test). The null and alternative hypothesis tested for MSR are

$$H_0: MSR = 0$$

$$H_1: MSR > 0.$$

The value of the MSR was compared to the critical value in a chi-square distribution with n-1 degrees of freedom to test whether the MSR is statistically different from 0 or not. If the MSR is statistically different from 0 then pairs of TIFs, TCCs, or SEs that are compared can be said to be statistically different from one another. Similarly, if the MSR is not statistically different from 0 then pairs of TIFs or TCCs that are compared can be said to be statistically similar to each other.

Results of the routing test and second-stage test analysis provided information on the effectiveness of the TST procedure for English- and French-speaking examinees who wrote the SAIP Science 1996 and 1999 tests. As discussed earlier, if the routing test has items that function differentially for English- and French-speaking examinees (identified by SIBTEST analysis of the routing test), then it can lead to misplacement of examinees in the second-stage test. Comparison of the English and French versions in the second-stage analysis either supports or rejects the above assumption.

CHAPTER IV: RESULTS

The School Achievement Indicators Program provides large samples and large numbers of items for examining the effect of translated test items on two-stage testing. The analyses were conducted in two steps. First, the routing test items were tested for DIF across English- and French-speaking examinees. Second, performance differences on the second-stage tests were compared for English- and French-speaking examinees by using different IRT based procedures discussed earlier in Chapter III.

Analysis of the Routing Test

Statistical Analysis

The routing test items were tested to identify items that function differentially for English- and French-speaking examinees. Summary statistics for the English and French test forms in each language for the 1996 and 1999 SAIP Science routing tests are presented in Table 4. As seen in Table 4, samples for the English and French forms for the 1996 and 1999 SAIP Science tests were large. Based on the mean score, it can be seen that for the 1996 and 1999 SAIP Science administrations, the ability differences between the French-speaking examinees and the English-speaking examinees within the 13- and 16-year-old groups were small. This is an advantage, since the more similar the groups, the more accurate the DIF detection (Hambleton et al. 1993).

The DIF analysis for the routing test was completed in two stages. At stage one, the 14 level 3 items common in both the easy and difficult second-stage tests were included in the DIF analysis to create a more stable matching subtest. Before including these items, DIF analysis including all 26 items (12 routing test items plus 14 common items) were performed. The DIF items identified in the set of common items were removed. The remaining items from this set were then combined with the 12 routing test items to form the final matching test for the second-stage analysis. The results of the DIF analysis for the 1996 and 1999 SAIP Science routing tests are summarized in Table 5. As seen in Table 5, for the 1996 SAIP Science test, three out of twelve items (items 1, 2, and 6) in the routing test were identified as DIF items for English- and French-speaking examinees ($\hat{B}_{UNI} > 0.059, p < 0.05$). Of these three items, two items favoured French-speaking examinees and one item favoured English-speaking examinees. For the 1999 SAIP Science test, the same three items (items 1, 2, and 6) in the routing test were identified as DIF items for English- and French-speaking examinees ($\hat{B}_{UNI} > 0.059, p < 0.05$). Of these three items, the same two items favoured French-speaking examinees and one item favoured English-speaking examinees.

As seen in the statistical analysis of the routing test, for the 1996 and 1999 administrations of SAIP Science, the same three items were identified as DIF across the two years. Further, the direction of DIF was also the same across the two years.

Substantive Analysis

The main purpose of the substantive analysis was to identify translation errors in the English- and French versions of the SAIP Science test, which consequently may lead to DIF in the routing test items. Four translators were asked to categorize the sources of differential item performance according to the four sources of translation errors identified by Gierl and Khaliq (2001). They were also asked to create their own sources of translation errors if the sources identified by Gierl and Khaliq (2001) were deemed inadequate or insufficient.

The substantive results obtained from the four translators are presented in Table 6. As seen in Table 6, the translators predicted two (items 3 and 8) out of 12 items in the routing test to favour English-speaking examinees. The translation errors ascribed to these items were found by Gierl and Khaliq (2001). Item 3 was believed to favour English-speaking examinees due to differences in words and expressions inherent to the language or culture. Item 8 was believed to favour English-speaking examinees due to differences in words and expressions not inherent to the language or culture. These two items were tested as a bundle for the 13- and 16-year-old, English- and French-speaking examinees for the 1996 and 1999 Science administrations. Results of the item bundle analysis are shown in Table 7. Since each bundle had only one item, results were interpreted using the DIF guidelines suggested by Roussous and Stout (1996) for single item DIF analysis. As seen in Table 7, the bundle analysis for items 3 and 8 show no DIF for 13-year-old, English- and French-speaking examinees for the 1996 Science test and 13- and 16-year-old, English- and French-speaking examinees for the 1999 Science test. However, for 16-year-old examines who wrote the 1996 Science test, the bundle consisting of item 8 showed moderate DIF, and it favoured the English-speaking examinees.

The four translators identified other differences in the English and French versions of the routing test items. However, according to the four translators, these translation differences would not lead to performance differences for English- and French-speaking examinees. For example, the English version of an item in the routing test used the word "chewed" and the French version of the item used the word " *sont mange'es* ", which means eaten. In the English version of the item, "the leaves were merely chewed, not necessarily eaten." However, according to the translators, this difference would not lead to performance differences for English- and French-speaking examinees (e.g., if the difficulty level of the item is not altered by this translation difference, then English- and French-speaking examinees may perform equally well on this item even in the presence of translation differences).

As noted in the statistical analysis of the routing test, there were three items that were consistently identified as displaying DIF for English- and French-speaking examinees for both the 1996 and 1999 SAIP Science tests. However, the four translators failed to identify these three items as problematic in their substantive review. They also reported that there was no translation error associated with these items that could possibly lead to performance differences between English- and French speaking examinees. Hence, the source for differential item performance on these items could not be identified.

Analysis of the Second-Stage test

Results of the second-stage tests for the SAIP Science 1996 and 1999 examinations are presented under six major headings discussed earlier in Chapter III. The results for the 1996 Science test are presented first followed by the results of the 1999 Science test.

Results of the 1996 SAIP Science Test

Estimated Latent Distributions: 1996 Administration

The latent distributions estimated for English- and French-speaking examinees within low- and high-ability groups for the second-stage tests are depicted in Figures 10 and 11. These latent distributions have a mean of zero and standard deviation of one. Figure 10 shows the latent distributions for 13-year-old, low ability, English- and French-speaking examinees and 13-year-old, high ability, English- and French-speaking examinees. Similarly, Figure 11 shows the latent distributions for 16-year-old, low ability, English- and French-speaking examinees and 16-year-old, high ability, English- and French-speaking examinees. As seen in Figures 10 and 11, the latent distributions for the low ability examinees are shifted to the left of the proficiency scale and the latent distributions for the high ability examinees are shifted to the right of the proficiency scale. The latent distributions show that the latent trait (e.g., science ability) for the English- and French-speaking examinees in the low- and high-ability groups are very similar which is evident by the large overlap between the two distributions. The mean and standard deviation (SD) for the latent distributions are also shown in Figures 10 and 11. The means of the latent distributions for the low-ability English- and French-speaking examinees are negative. In contrast, the means for the latent distributions for the high-ability English- and French-speaking examinees are positive. These similarities between the latent distributions for English- and French-speaking examinees in the low- and high-ability groups suggests that the English- and French-speaking examinees were placed in a similar location on the score scale.

TIFs for English and French Versions of the Tests: 1996 Administration

The TIFs for the second-stage tests are depicted in Figures 12 and 13. Figure 12 shows the TIFs for the English and French versions of the test for 13-year-old, low and high ability examinees. Similarly, Figure 13 shows the TIFs for the English and French versions of the test for 16-year-old, low and high ability examinees. As seen in Figure 12, the TIFs for the English and French versions of the test appears to be different for both low- and high ability, 13-year-olds. Similarly, as seen in Figure 13, the TIFs for the English and French versions of the test also appear to be different for low- and high-ability ability, 16-year-olds. As seen in Figure 12, for the 13-year- old, high ability examinees, the French version of the test gives slightly more information than the English version of the test. Also, as seen in Figure 13, for the 16-year-old, low and high ability examinees, the French version of the test gives slightly more information as than the English version of the test.

As discussed earlier, if examinees are not routed properly, then there can be misplacement of examinees in the second-stage test (e.g., high ability French speaking examinees taking an easy test). This, in turn, will result in the estimation of information for the English and French versions of the second-stage test for English- and French-speaking examinees that are different from one another. The results might indicate that both low-and high-ability English- and French-speaking examinees were not routed properly as the TIFs are less comparable to each other.

TCCs for English and French Versions of the Tests: 1996 Administration

The TCCs for the second-stage tests are shown in Figure 14 (Panels A, B, C & D). Panels A and B show the TCCs for the English and French versions of the test for

13-year-old, low and high ability examinees, respectively. Similarly, Panels C and D show the TCCs for the English and French versions of the test for 16-year-old, low and high ability examinees, respectively. As seen in Panels A, B, C and D, the TCCs for the English and French versions of the test appears to be slightly different for both low- and high-ability ability, 13-and 16-year-old examinees.

These results might indicate that both low-and high-ability English- and French-speaking examinees were not routed properly as the TCCs are less comparable to each other.

RE for English and French Versions of the Tests: 1996 Administration

The relative efficiency functions (RE) for the second-stage tests are shown in Figure 15 (Panels A, B, C, and D). Panels A and B show the RE for the English and French versions of the test for 13-year-old, low and high ability examinees, respectively. Similarly, Panels C and D shows the RE for the English and French versions of the test for 16-year-old, low and high ability examinees, respectively. As seen in Panels A and B, the RE for the French and English version suggests that the French version yields a better precision of measurement than the English version at the lower end of the ability scale but the English version yields a better precision of measurement than the French version at the higher end of the ability scale. Similarly, as seen in Panels C and D, the RE for the French and English version suggests that the French version of the test yields a better precision of measurement than the English version at the lower end of the ability scale but the English version yields a better precision of measurement than the French version at the higher end of the ability scale.

SE ($\hat{\theta}$) for English and French versions of the Tests: 1996 Administration

The SE ($\hat{\theta}$) for the second-stage tests are depicted in Figures 16 and 17. Figure 16 shows the *SE ($\hat{\theta}$)* for the English and French versions of the test for 13-year-old, low and high ability examinees. Similarly, Figure 17 shows the *SE ($\hat{\theta}$)* for the English and French versions of the test for 16-year-old, high and low ability examinees. As seen in Figure 16, the SE ($\hat{\theta}$) for the English and French versions of the test appears to be slightly different for both low- and high-ability, 13-year-olds. Similarly, SE ($\hat{\theta}$) for the English and French versions of the test appears to be slightly different for both low- and high-ability, 16-year-olds. Furthermore, as seen in Figures 16 and 17, the SE ($\hat{\theta}$) for low ability, 13-and 16-year-olds is higher at the higher end of the ability scale. However, the SE ($\hat{\theta}$) for high ability 13- and 16-year-olds is higher at the lower end of the ability scale.

Reliability Indices for English and French versions of the Tests: 1996 Administration

The reliability indices calculated for the second-stage tests are shown in Table 8. As discussed earlier, reliability indices between the range of 0.8 to 0.9 are considered appropriate for low-stakes examinations such as conducted by SAIP. As seen in Table 8, this range of reliability is achieved for the English and French versions of the tests for English- and French-speaking examinees in both low- and high-ability groups. Furthermore, similarity in the reliability indices for the English- and French-speaking examinees within the low- and high- ability groups may also suggest that English- and French- speaking examinees within the low- and high-ability groups performed similarly in the second-stage test.

The results of the comparisons of the TIFs, TCCs and SE(θ) for English and French versions of the second-stage tests using MSR are shown in Table 9. Analysis of the second-stage tests show that although graphical approaches suggested that the TIFs, TCCs, and SE ($\hat{\theta}$) show differences for French and English versions of the tests for both low- and high-ability examinees, results of the MSR failed to statistically support the findings obtained using the graphical approach. As seen in Table 9, the MSR values for all the comparisons are not statistically significant ($p > 0.05$) suggesting that the TIFs, TCCs, and SE ($\hat{\theta}$) were statistically similar for English and French versions of the tests for both 13- and 16-year-old, low and high ability examinees.

Results of the 1999 SAIP Science Test

Estimated Latent Distributions: 1999 Administration

The latent distributions estimated for English- and French-speaking examinees within low- and high-ability groups for the second-stage tests are depicted in Figures 18 and 19. These latent distributions have a mean of zero and standard deviation (σ) of one. Figure 18 shows the latent distributions for 13-year-old, low ability, English- and French-speaking examinees and 13-year-old, high ability, English- and French-speaking examinees. Similarly, Figure 19 shows the latent distributions for 16-year-old, low ability, English- and French-speaking examinees and 16-year-old, high ability, English- and French-speaking examinees. As seen in Figures 18 and 19, the latent distributions for the low ability examinees are shifted to the left of the proficiency scale and the latent distributions for the high ability examinees are shifted to the right of the proficiency scale. The latent distributions show that the latent trait (e.g., science ability) for both the English- and French-speaking examinees in the low- and high-ability groups are very similar which is evident by the large overlap between the two distributions. The mean and standard deviation (SD) for the latent distributions are also shown in Figures 18 and 19. The means for the latent distributions for the low-ability English- and French-speaking examinees are negative. In contrast, the means for the latent distributions for the high-ability English- and French-speaking examinees are positive. These similarities between the latent distributions for English- and French-speaking examinees within the low- and high-ability groups suggests that the English- and French-speaking examinees were placed in a similar location on the score scale.

TIFs for English and French Versions of the Tests: 1999 Administration

The TIFs for the second-stage tests are depicted in Figures 20 and 21. Figure 20 shows the TIFs for the English and French versions of the test for 13-year-old, low and high ability examinees. Similarly Figure 21, shows the TIFs for the English and French versions of the test for 16-year-old, low and high ability examinees. As seen in Figure 20, the TIFs for the English and French versions of the test appears to be more different for low ability, 13-year-olds than high ability, 13-year-olds in which, the TIFs are more similar to each other. Similarly, as seen in Figure 21, the TIFs for the English and French versions of the test appears to be more different for low ability, 16-year-olds than high ability, 16-year-olds in which, the TIFs are more similar to each other. Also as seen in Figure 20, for 13-year- old, low ability examinees, the English version of the test gives slightly more information than the

French version of the test. However, as seen in Figure 21, for the 16-year-old, low ability examinees, the French version of the test gives slightly more information as than the English version of the test.

As discussed earlier, if examinees are not routed properly, then there can be misplacement of examinees in the second-stage test (e.g., high-ability French speaking examinees taking an easy test). This, in turn, will result in the estimation of information for the English and French versions of the second-stage test for English- and French-speaking examinees that are different from one another. These results might indicate that the English- and French-speaking examinees were not routed properly to the low-ability group as the TIFs are less similar to each other. However, the English- and French-speaking examinees may be properly routed to the high-ability group as the TIFs are more similar to each other.

TCCs for English and French Versions of the Tests: 1999 Administration

The TCCs for the second-stage tests are shown in Figure 22 (Panels A, B, C & D). Panels A and B show the TCCs for the English and French versions of the test for 13-year-old, low and high ability examinees, respectively. Similarly, Panels C and D show the TCCs for the English and French versions of the test for 16-year-old, low and high ability examinees, respectively. As seen in Panels A and C, the TCCs for the English and French versions of the test appears to be different for low ability, 13-and 16-year-olds. However, as seen in Panels B and D, the TCCs for the English and French versions of the test appears to be similar for high-ability, 13- and 16-year-olds.

These results might indicate that the low-ability English- and French-speaking examinees were not routed properly as the TCCs are less similar to each other. However, for the high-ability English- and French-speaking examinees the TCCs are more similar to each other suggesting that there was proper routing in the first-stage test.

RE for English and French Versions of the Tests: 1999 Administration

The relative efficiency functions (RE) for the second-stage tests are shown in Figure 23 (Panels A, B, C, and D). Panels A and B show the RE for the English and French versions of the test for 13-year-old, low and high ability examinees, respectively. Similarly, Panels C and D shows the RE for the English and French versions of the test for 16-year-old, low and high ability examinees, respectively. As seen in Panels A and C, the RE for the French and English version suggests that the French version yields a better precision of measurement than the English version at the lower end of the ability scale but the English version yields a better precision of measurement than the French version at the higher end of the ability scale. However, as seen in Panels B and D, the RE for the French and English version suggests that the French version of the test yields a comparable precision of measurement as the English version at all points on the ability scale.

SE ($\hat{\theta}$) for English and French Versions of the Tests: 1999 Administration

The SE ($\hat{\theta}$) for the second-stage tests are depicted in Figures 24 and 25. Figure 24 shows the *SE* ($\hat{\theta}$) for the English and French versions of the test for 13-year-old, low and high ability examinees. Similarly Figure 25, shows the *SE* ($\hat{\theta}$) for the English and French versions of the test for 16-year-old, high and low ability examinees. As seen in Figure 24, the SE ($\hat{\theta}$) for the English and French versions of the test appears to be more different for low ability, 13-year-olds than high ability, 13-year-olds. Similarly, as seen in Figure 25, the SE ($\hat{\theta}$) for the English and French versions of the test appears to be more different for low ability, 16-year-olds than high ability, 16-year-olds. Further, as seen in Figures 24 and 25, the SE ($\hat{\theta}$) for low ability, 13-and 16-year-olds is higher at the higher end of the ability scale. However, the SE ($\hat{\theta}$) for high ability 13- and 16-year-olds is higher at the lower end of the ability scale.

Reliability Indices for English and French Versions of the Tests: 1999 Administration

The reliability indices calculated for the second-stage tests are shown in Table 10. According to Bock and Zimowski (1998), reliability indices between the range of 0.8 to 0.9 are considered appropriate for low-stakes examinations such as conducted by SAIP. As seen in Table 10, this range of reliability is achieved for the English and French versions of the tests for English- and French-speaking examinees in both low- and high-ability groups. Furthermore, similarity in the reliability indices for the English- and French-speaking examinees within the low- and high- ability groups may also suggest that English- and French- speaking examinees with the low- and high- ability groups performed similarly in the second-stage test.

The results of the comparisons of the TIFs, TCCs, and SE ($\hat{\theta}$) for English and French versions of the second-stage tests using MSR are shown in Table 11. Although, graphical approaches suggested that the TIFs, TCCs, and SE ($\hat{\theta}$) were more different for French and English versions of the tests for low ability examinees than for high-ability examinees, results of the MSR failed to substantiate the findings. As seen in Table 11, the MSR values for all the comparisons are not statistically significant ($p > 0.05$) suggesting that the TIFs, TCCs, and SE ($\hat{\theta}$) were statistically similar for English and French versions of the tests for both 13- and 16-year-old, low and high ability examinees.

In summary, results from the SAIP Science 1996 and 1999 analyses provide information on the effectiveness of the TST procedure for English- and French-speaking examinees. If the routing test has items that function differentially for English- and French-speaking examinees, then it can lead to misplacement of examinees in the second-stage test. Results of the statistical analysis of the routing test suggested that there were three out of twelve items that showed DIF for English- and French-speaking examinees. However, substantive analysis of the routing test items failed to identify the sources of differential performance for these three items. The four translators did not identify any translation error for these items that would have lead to performance differences between English- and French speaking examinees. The translators predicted two items in the routing test to favour English-speaking examinees. However, when these items were tested as a bundle, only one bundle displayed DIF and the bundle favoured English-speaking examinees. Since the bundle displayed DIF for only one comparison, this outcome does not appear to be systematic.

Results from the second-stage analyses suggest that the routing test properly placed English- and French-speaking examinees to the second-stage test. As evident by the comparison of the test information functions, test characteristic curves, standard errors of estimates, relative efficiency curves and reliability indices, English- and French-speaking examinees did not show performance differences in the second-stage test.

CHAPTER V: DISCUSSION AND CONCLUSIONS

This chapter is organized in four sections. In the first section, the research questions and a brief description of the methods used in the present study are presented. A summary and discussion of the key findings are presented in the second section. The limitations of the study are presented in the third section. The last section contains the implications for practice and recommendations for future research.

Summary of Research Questions and Methods

The primary purpose of this study was to evaluate the effectiveness of a two-stage testing (TST) procedure for English- and French-speaking examinees who wrote the SAIP Science 1996 and 1999 examinations. Two-stage testing (TST) is an adaptive testing procedure where different test forms of varying difficulty are administered to examinees based on previous test performance. The key element in effective TST is the accuracy of the routing test. If the routing test does not place different groups of examinees equally well to each test form, then examinees can be assigned to an inappropriate second-stage test. The School Achievement Indicators Program (SAIP), which is the national achievement test in Canada, uses a TST procedure to assess educational progress of 13- and 16-year-olds in Science. These examinations are also administered in English and in French, Canada's two official languages. SAIP works with the implicit assumption that the routing test works equally well for examinees in English and French. If this assumption is true, then there should be proper placement of English- and French-speaking examinees in the second-stage test which, in turn, will lead to reliable and valid comparisons between English- and French-speaking examinees on the second-stage test. However if the assumption is not true, then there might be misplacement of English- and French-speaking examinees in the second-stage test. For example, if the routing test has items that are biased and favor French-speaking examinees, then more French-speaking examinees will be routed to a high-ability group even though they should have been routed to a low-ability group. Consequently, the French-speaking examinees will write a more difficult second-stage test and may perform poorly compared to the English-speaking examinees. Such an outcome may be considered unfair because it could adversely affect the reported achievement levels of French-speaking examinees. Since fairness is an important concern in the field of educational measurement, it is important to ensure that a particular form of testing, such as TST, does not unfairly favor one language group compared to another. Hence, the present study compared the performance of English and French examinees that wrote the SAIP Science 1996 and 1999 achievement test administered using the TST procedure.

More specifically, the following questions were addressed in this study:

1. Is there evidence of differential item performance for English- and French-speaking examinees on the routing test used in the TST procedure?
2. If so, what is the source of differential item performance in the routing test?
3. Is there evidence of differential test performance for English- and French-speaking examinees on the second-stage test used in the TST procedure?
4. Is there a relationship between performance on the routing test and the second-stage test used in the TST procedure?

To answer the *first* question, statistical analyses for the routing test of the SAIP Science 1996 and 1999 tests were conducted using SIBTEST to identify items that function differentially for English- and French-speaking examinees. The guidelines suggested by Roussos and Stout (1996) were used to classify DIF items in the present study. Items with a B- or C-level rating were considered DIF items whereas those with an A-level rating were not considered DIF items.

To answer the *second* question, substantive analysis was conducted to identify *why* items functioned differentially between groups. A translation review process developed by Gierl and Khaliq (2001) was used in the present study. In the translation review process, the translators were asked to evaluate the similarities and differences of the English and French test items by comparing the items across languages in the content area of Science. In each case, the translators were asked to specify which language group would be favored and also to identify the exact reason or reasons for the difference in each item. Once the task was completed, a group discussion and consensus for rating translation differences between the two language versions of the items was conducted.

To answer the *third* question, performance differences between English- and French-speaking examinees were compared using IRT based procedures. First, test information functions (TIFs) were compared for English and French versions of the test. Second, tests characteristic curves (TCCs) were compared for English and French versions of the test. Third, relative efficiency of the French compared with the English versions of the test was computed. Fourth, standard errors of estimate, denoted as $SE(\hat{\theta})$, were compared for English and French versions of the test. Fifth, reliability indices were calculated for the English and French versions of the test to obtain an estimate of the effectiveness of the TST procedure in general.

To answer the *fourth* question, the performance of English- and French-speaking examinees on the routing test and the second-stage test was examined. If the routing test has items that are free from the effects of DIF, then English- and French-speaking examinees should perform equally well on the routing test and, hence, English- and French-speaking examinees of equal ability should be assigned to the same second-stage test. Results of the second-stage analysis, where performance of English- and French-speaking examinees are compared, will either support or refute this assumption.

Findings

The results of the study indicated that English- and French-speaking examinees performed equally well on the second-stage tests thereby suggesting that English- and French-speaking examinees were properly placed by the routing test.

Routing Test Analyses

Statistical analysis of the SAIP Science 1996 routing test revealed three out of twelve items functioned differentially for English- and French-speaking examinees. Two of the items favoured French-speaking examinees and one item favoured English-speaking examinees. This finding was replicated in the SAIP Science 1999 routing test. However, this statistical outcome does not necessarily suggest that the three items were biased for English- or French-speaking examinees. As discussed in Chapter II, if the difference in performance for two different groups on a particular item is the result of actual differences in ability, then the item is not considered as biased. Instead the item may display impact which is not a negative attribute of an item.

To understand the nature of differences for the items displaying DIF, two analyses were conducted. First, the DIF items were compared to the substantive reviews. In this analysis, the four translators did not identify any translation errors for the DIF items flagged by SIBTEST. This finding leads to two possible conclusions. First, there were no translation errors associated with these items. Therefore, the differences in performance identified by the statistical analysis may reflect actual differences in ability between English and French examinees (i.e., impact). Since impact is not considered a negative attribute, the items identified as DIF may be considered as valid test items. Second, the translators were asked to focus on translation errors, not actual group differences. Hence, the translators may have failed to identify sources of translation errors that can lead to performance differences between English- and French-speaking examinees in which case the items may be considered biased. Hence, further studies need to be conducted on the routing test items to identify the sources of performance difference----both attributable to impact and bias----for the items identified as DIF using SIBTEST.

Second, a bundle analysis was conducted where items were first grouped by the test translators and then tested statistically. In this analysis, the statistical test did not identify the item bundles as displaying DIF in the English versus French comparisons for 13- and 16-year-old examinees who wrote the 1999 SAIP Science test. This result was also found for 13-year-old examinees who wrote the 1996 SAIP Science test. However, for the 16-year-old examinees who wrote the 1996 Science test, one bundle displayed DIF and the bundle favoured English-speaking examinees. Since the bundle displayed DIF for only one comparison, this outcome does not appear to be systematic. In other words, if the item does have a translation error, this error only affects one of the four group comparisons in this study. Further studies should be conducted on the routing test items by using different samples to evaluate whether the present finding can be replicated.

Second-Stage Test Analysis

Findings from Graphical Procedures

Graphical analysis of the second-stage tests for the SAIP Science 1996 test suggested that the test information functions (TIFs), test characteristic curves (TCCs), and standard errors or $SE(\theta)$ were less similar for English- and French-speaking examinees for both low- and high-ability groups. Relative efficiency for the French version compared with the English version for high-ability examinees suggested that the French version of the test yielded a higher measurement precision at the lower end of the ability scale. In contrast, the English version showed higher measurement precision at the higher end of the ability scale. Similarly, for low ability examinees, the relative efficiency of the French version compared to the English version suggested that the French version yielded more precision at the lower end of the ability scale but the English version yielded more precision at the higher end of the ability scale.

Graphical analysis of the second-stage tests for the SAIP Science 1999 test suggested that the test information functions (TIFs), test characteristic curves (TCCs), and standard errors or $SE(\theta)$ were more discrepant for English- and French-speaking examinees in the low ability group but less different for English- and French-speaking examinees in the high ability groups. Similarly, the relative efficiency for the French versions compared with the English version for high-ability examinees suggested that the French version of the test yielded a more similar precision of measurement as the English version at all points on the ability scale. However, the relative efficiency for the French version compared with the English version for low-ability examinees suggested that the French version yielded a better precision of measurement than the English version at the lower end of the ability scale. In contrast, the English version yielded more measurement precision than the French version at the higher end of the ability scale. This finding may be attributed to the fact that the routing test placed the high ability examinees in the second-stage-tests more accurately than the low-ability examinees.

The findings of the graphical procedure provided an estimate of the magnitude of difference between pairs of TIFs, TCCs, SEs, and REs. To get more precise estimates of the magnitude of difference between pairs of TIFs, TCCs, SEs, and REs, a statistical approach was used. Findings of the statistical method are discussed below.

Findings from Statistical Procedure

Although graphical approaches suggested that the TIFs, TCCs, and SEs were slightly different for English and French versions of the tests for both low- and high-ability examinees in the SAIP 1996 test, results of the mean square residual failed to support the findings. Similarly, graphical approaches suggested that the TIFs, TCCs, and SEs were different for English and French versions of the tests for low ability examinees but not for high-ability examinees in the SAIP 1999 test. Again, results of the mean square residual failed to support the findings. The results of the mean square residual comparisons suggested that all the TIF, TCC, SEs comparisons for the eight sub-groups studied, were not statistically significant ($p > 0.05$) suggesting that the TIFs TCCs, and SEs were statistically similar for English and French versions of the tests for both 13- and 16-year-old, low and high ability examinees.

Further, reliability indices for the English and French versions of the tests for the SAIP Science 1996 and 1999 tests were high for both 13- and 16-year-old, low- and high ability examinees. This outcome suggests that in general, the second-stage-tests worked equally well for English- and French-speaking examinees.

To conclude, statistical analysis of the routing test items suggested that only three out of twelve items displayed DIF for English- and French-speaking examinees. Further, substantive analysis of the routing test items revealed that translation errors were not the cause of DIF for the three items identified as DIF suggesting that the effects of negative DIF or bias on these items was minimal. Since the majority of the items in the routing test are free from DIF, it is reasonable to assume that the routing test placed the English- and French-speaking examinees equally well in the second-stage tests. This assumption was more strongly established by the findings from the second-stage test analysis where English- and French-speaking examinees performed equally well on the second-stage test. Although some discrepancies between pairs of TIFs, TCCs, and SEs were observed using the graphical method, failure to detect such differences using a statistical method suggests that such differences are likely attributable to random sampling error.

Limitations of the Study

The primary purpose of this study was to evaluate, in a psychometric sense, the effectiveness of a two-stage testing procedure for English- and French-speaking examinees who wrote the SAIP Science 1996 and 1999 tests. As discussed earlier, the routing test is the key element in two-stage testing. A routing test consisting of items with DIF can lead to improper estimation of examinees' ability which, consequently, might lead to improper placement of examinees in the second-stage test. In the present study DIF items in the routing test were identified using statistical as well as substantive methods. Substantive analyses included using test translators to identify sources of translation errors that could contribute to performance differences among English- and French-speaking examinees. However, several factors that were not considered in this study could be potential explanations for DIF. For example, different instructional methods could lead to group differences. Identifying these alternative factors may contribute to a better understanding about the nature of DIF on the routing test, which may lead to better decisions about what to do with items that are identified as DIF using statistical procedures. Also, little attention in this study was given to the actual cognitive processes that might be used by examinees as they respond to achievement test items in different languages. Protocol analysis (Ericsson & Simon, 1993) holds promise for helping researchers come to a better understanding of how the items are solved by the examinees. However, this method is both resource and time intensive. Hence, for the present study, it was not possible to use the method.

Another limitation of the study concerns the sample of students examined. It was not possible to completely isolate Francophones from French-immersion students. Francophones are students whose mother tongue is French and who speak, read, and write in French. French-immersion students are composed of students whose native tongue is not French, but who are enrolled in a French immersion school and read and write French in school. However, they may speak a different language at home. Hence, even though Francophones and French-immersion students may write a French version of a test, their problem-solving skills may be considerably different from one another. This sample characteristic could make interpretations from a substantive analysis less meaningful because differences in the English- and French-speaking groups may be due to systematic differences between English-speaking examinees and Francophones, or English-speaking examinees and French immersion students, or a combination of both. In the data used in the present study it was not possible to isolate the Francophones from the French immersion students in the 1999 data because there was no variable that could be used to partition the two groups. In the 1996 data it was possible to partition the Francophones and the French-immersion students. However, to have similar samples (for generalizability of results) across SAIP Science 1999 and 1996, it was decided not to partition the Francophones and French immersion students in this study.

Recommendations

Future Practice

The routing test is the most important element in two-stage testing. Hence, future research should be directed to a more thorough analysis of the routing test where many possible causes of differential performance between English- and French-speaking examinees are assessed. Since many educational tests (e.g., tests administered by the School Achievement Indicators Program) are based on cognitive problem-solving skills, it is necessary that cognitive components of test performance be understood to validly interpret results from educational tests (e.g., Frederiksen, Mislevy, & Bejar, 1993; Gierl, 1997; Hattie, Jaeger, & Bond, 1999). Therefore, apart from test translators, cognitive psychologists should also participate in the substantive analysis of the routing test to identify cognitive skills needed to solve specific test problems. The importance of identifying cognitive skills for understanding differences in test performance for different language groups such as English and French is evident in the following example. Substantive reviews by test translators may identify words or sentences that are different for English and French versions of an item thereby leading to the conclusion that these two languages will function differentially for English- and French speaking examinees. However, such a conclusion may be erroneous if the differences in words and sentences between the two language versions does not affect the difficulty level of the item or change the cognitive skills needed to solve the problem. In this case, performance differences may not occur between language or cultural groups even in the presence of translation differences. Hence, future research should be directed to a more comprehensive understanding of students' cognitive processes as they respond to test items in multiple languages.

Future Research

Two real data sets were used in the present study to evaluate the effectiveness of a TST procedure for English- and French-speaking examinees. Equivalence of the second-stage tests was assessed by comparing the test information functions, test characteristic curves, and standard errors of estimate for the English and French versions of the tests. Further, the relative efficiency of the French version of the tests was compared to the English version of the tests. The study was conducted based on the assumption that the second-stage tests are free from the effects of DIF. Consequently, performance differences on the second-stage tests can only be attributed to improper routing of examinees in the routing test.

However, results from the analysis may become difficult to interpret if there are DIF items present in the second-stage tests. The difficulty may arise due to the additional complexity of determining whether the differences in performance between English- and French-speaking examinees are a result of ability differences or the presence of biased items in the second-stage tests or a combination of both. Therefore, in the next stage in my program of research, purification of the second-stage tests using DIF analyses will be conducted to identify items that function differentially for English- and French-speaking examinees. By having a purer second-stage test, performance differences between English- and French-speaking examinees can be attributed to ability differences, which may result from improper routing in the routing test.

Purification of the second-stage tests would involve two stages. First, statistical analysis can be conducted to identify items that function differentially for

English- and French-speaking examinees. Use of multiple statistical methods can lead to converging evidence about DIF for different test items. Hence, apart from using SIBTEST for DIF analysis, other statistical tests will be used for assessing comparability among different language versions of test items. IRT-based methods for monitoring the comparability of test items for different groups of examinees can provide different information about the nature of DIF as compared to classical test theory based methods (e.g., SIBTEST). IRT methods allow DIF detection on different locations on the ability scale otherwise known as local DIF (Bolt & Gierl, 2002). This approach is not possible using classical test theory methods where groups of examinees are matched on the total test score. The detection of DIF at different points on the ability scale can provide valuable information to test developers and test translators regarding the nature of DIF. For example, if test translators have prior information that a particular item displays considerable amount of DIF on the lower end of the ability scale but not on the higher end of the ability scale, then test translators can use this information to focus on more specific causes of DIF. Hence, DIF will be identified using multiple methods.

Second, to understand the nature of DIF items identified by the statistical analysis, substantive analysis regarding the cause of DIF will be conducted. Sources of DIF on translated test items will be identified by using bilingual teachers (as was done in the current study), content specialists, and cognitive psychologists. Gierl and Khaliq (2001) emphasized the use of cognitive psychologists for detecting causes of DIF because cognitive problem-solving skills are an important component of most educational tests, yet little is known about the cognitive processes actually used by examinees as they respond to achievement test items in different languages. By using cognitive psychologists, content specialists, and bilingual teachers it will be possible to better understand the causes of DIF and also identify more sources of translation errors. The advantage of using content specialists include their familiarity of the content, the achievement tests and the language in which these tests are administered. Moreover by virtue of their experiences in teaching, the teacher will be familiar with the strategies typically used by students to solve problems on achievement tests. Finally, cognitive psychologists are familiar with the actual problem solving skills of examinees by virtue of their research on cognitive processing skills. Hence, they can provide information about the ways examinees solve various problems on translated tests from both a theoretical and applied perspective.

Again, protocol analysis (Ericcson & Simon, 1993) can contribute to our understanding of examinees' cognitive processes as they respond to test items. Since protocol analysis requires examinees to think aloud as they solve test problems, results from protocol analysis can provide first hand information about the ways examinees solve various problems on translated tests. At this stage, results from the substantive reviewers can be compared with the verbal protocols of the examinees which, in turn, will provide convergent validity evidence about the nature of cognitive processing skills of examinees and the sources of differential item performance on translated tests.

References

Ackerman, T. (1987). *The use of unidimensional item parameter estimates of multidimensional items in adaptive testing* (Research Rep. No. 87-13). Iowa City, IA: American College Testing Program.

Ackerman, T. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13,* 113-127.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9,* 37-48.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Allalouf, A., Hambleton, R., & Sireci, S. (1999). Identifying the causes of translation DIF on verbal items. *Journal of Educational Measurement, 36,* 185-198.

Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the prueba de aptitud academica and the scholastic aptitude test* (College Board Report No. 88-2). New York: College Entrance Examination Board.

Bock, R.D., & Zimowski, M.F. (1997). Multiple group IRT. In Van der Linden, W. J., & Hambleton, R.K (Eds.), *Handbook of modern item response theory* (pp. 433-448). New York: Springer.

Bock, R.D., & Zimowski, M.F. (1998). *Feasibility studies of two-stage testing in large-scale educational assessment: Implications for NAEP*. American Institutes for Research, CA.

Bolt, D. M., & Gierl, M. J. (2002). *Applications of a regression correction to three nonparametric tests of DIF: Implications for global and local DIF detection.* Manuscript submitted for publication.

Bolt, D., and Stout, W. (1996). Differential Item Functioning: Its multidimensional model and resulting SIBTEST detection procedure. *.Behaviormetrika, 23,* 67-95.

Boughton, K.A. (2001). *The construction and evaluation of automated parallel forms and multiple parallel panels in computer-adaptive sequential testing.* Unpublished doctoral thesis, University of Alberta, Edmonton, Canada.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items.* Newbury Park: Sage.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31-44.

Council of Ministers of Education, Canada (2000). *Report on science assessment, School achievement indicators program (1999)*, Toronto, ON: Council of Ministers of Education, Canada.

Douglas, J., Roussos, L., and Stout, W. (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. *Journal of Educational Measurement, 33,* 465-484.

Ercikan, K., Gierl, M.J., McCreith, T., Puhan, G., & Koh, K. (2002). *Comparability of English and French Versions of SAIP for reading, mathematics and science Items*. Paper presented at the annual meeting of the Canadian Society for the Study of Education, Toronto.

Ericcson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Fidalgo, A.M. (1996a). Funcionamiento diferencial de los items [Differential item functioning]. In J. Muñiz (Ed.), *Psicometria* (pp. 371-455). Madrid, Spain: Universitas.

Frederiksen, N., Mislevy, R. J., Bejar, I. I. (1993). *Test theory for a new generation of tests*. Hillsdale, NJ: Erlbaum.

Gierl, M. J. (1997). Comparing the cognitive representations of test developers and students on a mathematics achievement test using Bloom's taxonomy. *Journal of Educational Research, 91,* 26-32.

Gierl, M. J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education, 25,* 280-296.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests. *Journal of Educational Measurement, 38,* 164-187.

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999). *Consistency between statistical procedures and content reviews for identifying translation DIF*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Quebec, Canada.

Hambleton, R.K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment, 9,* 57-68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests:A progress report. *European Journal of Psychological Assessment, 10,* 229-224.

Hambleton, R.K. (2001). The next generation of ITCtest translation and adaptation guidelines. *European Journal of Psychological Assessment, 17,* 164-172.

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research, 45,* 153-171.

Hambleton, R.K., & Patsula, L. (1999). Increasing the Validity of Adapted Tests: Myths to be Avoided and Guidelines for Improving Test Adaptation Practices. *Journal of Applied Testing Technology, 1,* 1-30.

Hambleton, R.K., & Swaminathan, H. (1985*). Item response theory: Methods and practices*. New York: Springer.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991*). Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Jeanrie, C., & Bertrand, R. (1999). Translating tests with the International Test Commission's guidelines: Keeping validity in mind. *European Journal of Psychological Assessment, 15,* 277-283.

Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent methodological questions in educational testing. *Review of Research in Education, 24,* 393-446.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). *Item response theory: Application to psychological measurement*. Dow Jones-Irwin, IL: Homewood.

Jiang, H. & Stout, W. (1998). Improved Type I Error Control and Reduced Estimation Bias for DIF Detection Using SIBTEST. *Journal of Educational and Behavioural Statistics, 23 (4),* 291-322.

Joint Advisory Committee. (1993). *Principles for Fair Student Assessment Practices for Education in Canada.* Edmonton, AB: Centre for Research in Applied Measurement and Evaluation.

Lonner, W.J. (1990). An overview of cross-cultural testing and assessment. In R.W. Brislin (Ed.), *Applied cross-cultural psychology* (pp. 56-76). Newbury Park, CA: Sage Publications.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Messick, S. (1990). Validity. In R.L. Linn (Ed.), *Educational Measurement, Third edition* (pp. 13-103). American Council on Education and Macmillan Publishing Company.

Millsap, R. E. & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297-334.

Murphy, K.R. & Davidshofer, C.O. (1998). *Psychological Testing: Principles and applications*. Upper Saddle River, NJ: Prentice Hall, NJ.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28,* 99-117.

Ndalichako, J.L. & Rogers, W.T. (1997). Comparison of finite state score theory, classical test theory, and item response theory in scoring multiple-choice items. *Educational and Psychological Measurement, 57 (4),* 580-589.

Potenza, M. T. & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement*, 19, 23-37.

Reckase, M. D., & Kunce, C. (1999, May). *Translation accuracy of a technical credentialing examination*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Quebec, Canada.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement*, *33*, 215-230.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, *58*, 159-194.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of six procedures for detecting test item bias using both internal and external ability criteria. *Journal of Educational Statistics*, *6*, 317-375.

Sireci, S.G. (1996, April). *Technical issues in linking assessments across languages*. Paper presented at the annual meeting of the National Council on Educational Measurement, New York.

Sireci, S.G.(1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice, 16 (1),* 12-19.

Stout, W. (1990). A new item response modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrica, 55,* 293-325.

Thissen, D., Steinberg, L. & Gerrard, M. (1986). Beyond group mean differences: The concept of item bias. *Psychological Bulletin*, *99*, 118-128.

Van de Vijver, F & Leung, K (1997). *Methods and data analysis for cross-cultural research.* Thousand Oaks, CA: Sage.

Yen, W. M. (1984). Effects of local item independence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125-145.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Erlbaum.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Boch, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary Items*. Chicago: Scientific Software International.

Table 1.

Results of Item-Guessing Analysis for the Second-Stage Tests: 1996 SAIP Science Administration

	Low-ability Examinees											
	13-year olds						16-year olds					
	English			French			English			French		
Item No.	35	39	59	35	39	59	35	39	59	35	39	59
0	687	681	684	358	354	355	472	471	473	145	145	145
1	2	8	5	1	5	4	1	2	0	0	0	0
	High-ability Examinees											
Item No.	7	25	64	14	25	64	7	14	64	14	25	64
0	1267	1283	1278	589	593	594	765	755	767	346	347	348
1	16	0	5	5	1	0	2	12	0	2	1	0

Note. 0 = No. of examinees scoring zero on item *i*
1 = No. of examinees scoring one on item *i*

Table 2.

Results of Item-Guessing Analysis for Eight Sub-groups on the Second-Stage Tests; 1999 SAIP Science Administration

	Low-ability Examinees											
	13-year olds						16-year olds					
	English			French			English			French		
Item No.	15	59	64	15	27	59	24	59	64	15	27	59
0	395	401	344	191	174	197	90	104	89	47	38	47
1	10	4	61	8	25	2	17	3	18	1	10	1
	High-ability Examinees											
Item No.	25	40	41	25	41	64	25	40	41	25	41	64
0	840	831	842	468	469	461	379	375	382	224	227	224
1	10	19	8	3	2	10	5	9	2	3	0	3

Note. 0 = No. of examinees scoring zero on item *i* and 1 = No. of examinees scoring one on item *i*

Table 3.

Results of Test-Speededness Analysis for the Easy- and Difficult Second-Stage Tests: 1996 and 1999 SAIP Science Tests

	1996							
	Low-ability Examinees				High-ability Examinees			
	13-year Olds		16-year Olds		13-year Olds		16-year Olds	
	English	French	English	French	English	French	English	French
% of examinees completing 75% of the items	93.40	92.90	85.65	91.31	98.13	97.25	98.59	98.06
	1999							
% of examinees completing 75% of the items	88.21	90.77	82.11	87.60	94.63	95.16	96.95	97.41

Table 4.

Summary Statistics for the English and French Versions of the 1996 and 1999 SAIP Science Tests (Routing Test Analysis)

		1996		
	13-year Olds		16-year Olds	
	English	French	English	French
Sample Size	7000[a]	3526	7000	3113
No. of Items	12	12	12	12
Mean	12.71	12.23	15.27	14.93
SD	4.10	3.79	3.98	3.82
		1999		
Sample Size	7000	3008	7000	2898
No. of Items	12	12	12	12
Mean	14.24	13.87	17.80	17.24
SD	4.55	4.17	4.28	4.08

Note. Means for English- and French-speaking examinees within the 13- and 16-year old groups are significantly different at $p < 0.01$ in the t-test comparisons. However, due to the increased power resulting from the large sample sizes, these findings are misleading. Therefore the effect sizes were examined and they were close to zero suggesting that there are no important mean differences between the groups.

The mean scores for 13- and 16-year old examinees in the 1996 and 1999 SAIP Science tests are greater than the total number of items in the routing test (n=12) because the DIF analysis also included the common items from the easy and difficult second-stage tests (see p. 41).

[a] Random samples of 7000 examinees were selected for the English-speaking groups as it is the maximum sample size that SIBTEST allows for conducting DIF analysis.

Table 5.

Results of DIF Analysis using SIBTEST for the SAIP Science Routing Tests: 1996 and 1999 Administrations

	Favours	No. of DIF items	Item No.	$\hat{B}$ uni (Effect Size)	DIF level	*p*-value
			1996			
13-year Olds	French	2	1	-0.144	C	0.000
			2	-0.217	C	0.000
	English	1	6	0.171	C	0.000
16-year Olds	French	2	1	-0.087	B	0.000
			2	-0.170	C	0.000
	English	1	6	0.147	C	0.000
			1999			
13-year Olds	French	2	1	-0.154	C	0.000
			2	-0.225	C	0.000
	English	1	6	0.131	C	0.000
16-year Olds	French	2	1	-0.072	B	0.000
			2	-0.141	C	0.000
	English	1	6	0.111	C	0.000

Note. A negative $\hat{B}$ indicates the item favours French examinees
$p < .01$

Table 6.

Results of Substantive DIF Analysis for the SAIP Science Routing Test

Source	No. of Items	Item No.	Predicted to Favour
2 – Differences in Words or Expressions Inherent to Language or Culture	1	3	English
3 – Differences in Words or Expressions Not Inherent to Language or Culture	1	8	English
No Identifiable Sources of Translation Errors that can cause DIF	10	1, 2, 4, 5, 6, 7, 9, 10, 11, 12	None

Note. The items on the 1996 and 1999 SAIP Science Routing test were identical

Table 7.

Results of Bundle DIF Analysis using SIBTEST for the SAIP Science Routing Tests: 1996 and 1999 Administrations

			1996			
	No. of Items	Item No.	Predicted to Favour	$\hat{B}$ (Effect Size)	DIF level	*p*-value
13-year Olds	1	3	English	0.006	A	0.580
	1	8	English	0.025	A	0.017
16-year Olds	1	3	English	0.023	A	0.015
	1	8	English	0.063	B	0.000
			1999			
13-year Olds	1	3	English	0.020	A	0.056
	1	8	English	0.032	A	0.003
16-year Olds	1	3	English	0.008	A	0.422
	1	8	English	0.040	A	0.000

Note. A negative $\hat{B}$ indicates the item favours French examinees

* $p < .01$

Table 8.

Reliability Indices for English and French versions of the SAIP Science Second-Stage-Tests for Eight Groups of Examinees: 1996 Administration

	Low-ability Examinees				High-ability Examinees			
	13-year Olds		16-year Olds		13-year Olds		16-year Olds	
	English	French	English	French	English	French	English	French
Reliability Index	0.899	0.901	0.900	0.911	0.870	0.885	0.877	0.893

Table 9.

Results of the English and French Comparisons of the TIFs, TCCs, and SE(θ): 1996 SAIP Science Administration

	Low-ability Examinees		High-ability Examinees	
	13-year Olds	16-year Olds	13-year Olds	16-year Olds
	English/French[a]	English/French	English/French	English/French
Comparison of TIFs				
MSR	1.59	1.24	0.77	1.30
Comparison of TCCs				
MSR	0.30	0.90	0.58	0.98
Comparison of SE(θ)				
MSR	0.00	0.00	0.00	0.00

Note. Critical □^{2} at 30 degrees of freedom is 14.953. *$p < .05$

[a] Comparison of English and French versions of the second-stage tests.

Table 10.

Reliability Indices for English and French versions of the SAIP Science Second-Stage-Tests for Eight Groups of Examinees: 1999 Administration

	Low-ability Examinees				High-ability Examinees			
	13-year Olds		16-year Olds		13-year Olds		16-year Olds	
	English	French	English	French	English	French	English	French
Reliability Index	0.913	0.913	0.911	0.919	0.883	0.883	0.888	0.889

Table 11.

Results of the English and French Comparisons of the TIFs, TCCs, and SE(θ): 1999 SAIP Science Administration

		TIFs		
	Low-ability Examinees		High-ability Examinees	
	13-year Olds	16-year Olds	13-year Olds	16-year Olds
	English/French	English/French	English/French	English/French
MSR	2.27	3.23	0.07	0.26
		TCCs		
MSR	1.67	0.90	0.11	0.02
		SE(θ)		
MSR	0.00	0.00	0.00	0.00

Note. Critical χ^2 at 30 degrees of freedom is 14.953. *$p < .05$

[a]Comparison of English and French versions of the second-stage tests.

Figure Caption

Figure 1. Comparison of ICCs for English and French versions of a hypothetical achievement test.

Figure 2. Comparison of TCCs for English and French versions of a hypothetical achievement test.

Figure 3. Comparison of TIFs for English and French versions of a hypothetical achievement test.

Figure 4. Relative efficiency curve for English and French versions of a hypothetical achievement test.

Figure 5. Comparison of SE (θ) for English and French versions of a hypothetical achievement test.

Figure 6.

Panel A. Scree plot for 13-year-old, low-ability, English-speaking examinees:1996 administration.

Panel B. Scree plot for 13-year-old, low-ability, French-speaking examinees:1996 administration.

Panel C. Scree plot for 16-year-old, low-ability, English-speaking examinees: 1996 administration.

Panel D. Scree plot for 16-year-old, low-ability, French-speaking examinees:1996 administration.

Figure 7.

Panel A. Scree plot for 13-year-old, high-ability, English-speaking examinees: 1996 administration.

Panel B. Scree plot for 13-year-old, high-ability, French-speaking examinees: 1996 administration.

Panel C. Scree plot for 16-year-old, high-ability, English-speaking examinees: 1996 administration.

Panel D. Scree plot for 16-year-old, high-ability, English-speaking examinees: 1996 administration.

Figure 8.

Panel A. Scree plot for 13-year-old, low-ability, English-speaking examinees:1999 administration.

Panel B. Scree plot for 13-year-old, low-ability, French-speaking examinees:1999 administration.

Panel C. Scree plot for 16-year-old, low-ability, English-speaking examinees: 1999 administration.

Panel D. Scree plot for 16-year-old, low-ability, French-speaking examinees:1999 administration.

Figure 9.

Panel A. Scree plot for 13-year-old, high-ability, English-speaking examinees: 1999 administration.

Panel B. Scree plot for 13-year-old, high-ability, French-speaking examinees: 1999 administration.

Panel C. Scree plot for 16-year-old, high-ability, English-speaking examinees: 1999 administration.

Panel D. Scree plot for 16-year-old, high-ability, English-speaking examinees: 1999 administration.

Figure 10. Latent distributions for 13-year-old, low-ability, English- and French-speaking examinees and 13-year-old, high-ability, English- and French-speaking examinees: 1996 administration.

Figure 11. Latent distributions for 16-year-old, low-ability, English- and French-speaking examinees and 16-year-old, high-ability, English- and French-speaking examinees: 1996 administration.

Figure 12. TIFs for the English and French versions of the test for 13-year-old, low and high-ability examinees: 1996 administration.

Figure 13. TIFs for the English and French versions of the test for 16-year-old, low and high-ability examinees: 1996 administration.

Figure 14.

Panel A. TCCs for the English and French versions of the test for 13-year-old, low-ability examinees: 1996 administration.

Panel B. TCCs for the English and French versions of the test for 13-year-old, high-ability examinees: 1996 administration.

Panel C. TCCs for the English and French versions of the test for 16-year-old, low-ability examinees: 1996 administration.

Panel D. TCCs for the English and French versions of the test for 16-year-old, high-ability examinees: 1996 administration.

Figure 15.

Panel A. RE for the English and French versions of the test for 13-year-old, low-ability examinees: 1996 administration.

Panel B. RE for the English and French versions of the test for 13-year-old, high-ability examinees: 1996 administration.

Panel C. RE for the English and French versions of the test for 16-year-old, low-ability examinees: 1996 administration.

Panel D. RE for the English and French versions of the test for 16-year-old, high-ability examinees: 1996 administration.

Figure 16. SE ($\hat{\theta}$) for the English and French versions of the test for 13-year-old, low and high-ability examinees: 1996 administration.

Figure 17. SE ($\hat{\theta}$) for the English and French versions of the test for 16-year-old, high and low-ability examinees: 1996 administration.

Figure 18. Latent distributions for 13-year-old, low-ability, English- and French-speaking examinees and 13-year-old, high-ability, English- and French-speaking examinees: 1999 administration.

Figure 19. Latent distributions for 16-year-old, low-ability, English- and French-speaking examinees and 16-year-old, high-ability, English- and French-speaking examinees: 1999 administration.

Figure 20. TIFs for the English and French versions of the test for 13-year-old, low and high-ability examinees: 1999 administration.

Figure 21. TIFs for the English and French versions of the test for 16-year-old, low and high-ability examinees: 1999 administration.

Figure 22.

Panel A. TCCs for the English and French versions of the test for 13-year-old, low-ability examinees: 1999 administration.

Panel B. TCCs for the English and French versions of the test for 13-year-old, high-ability examinees: 1999 administration.

Panel C. TCCs for the English and French versions of the test for 16-year-old, low-ability examinees: 1999 administration.

Panel D. TCCs for the English and French versions of the test for 16-year-old, high-ability examinees: 1999 administration.

Figure 23.

Panel A. RE for the English and French versions of the test for 13-year-old, low-ability examinees: 1999 administration.

Panel B. RE for the English and French versions of the test for 13-year-old, high-ability examinees: 1999 administration.

Panel C. RE for the English and French versions of the test for 16-year-old, low-ability examinees: 1999 administration.

Panel D. RE for the English and French versions of the test for 16-year-old, high-ability examinees: 1999 administration.

Figure 24. SE $(\hat{\theta})$ for the English and French versions of the test for 13-year-old, low and high-ability examinees: 1999 administration.

Figure 25. SE $(\hat{\theta})$ for the English and French versions of the test for 16-year-old, high and low-ability examinees: 1999 administration.

Figure 1.

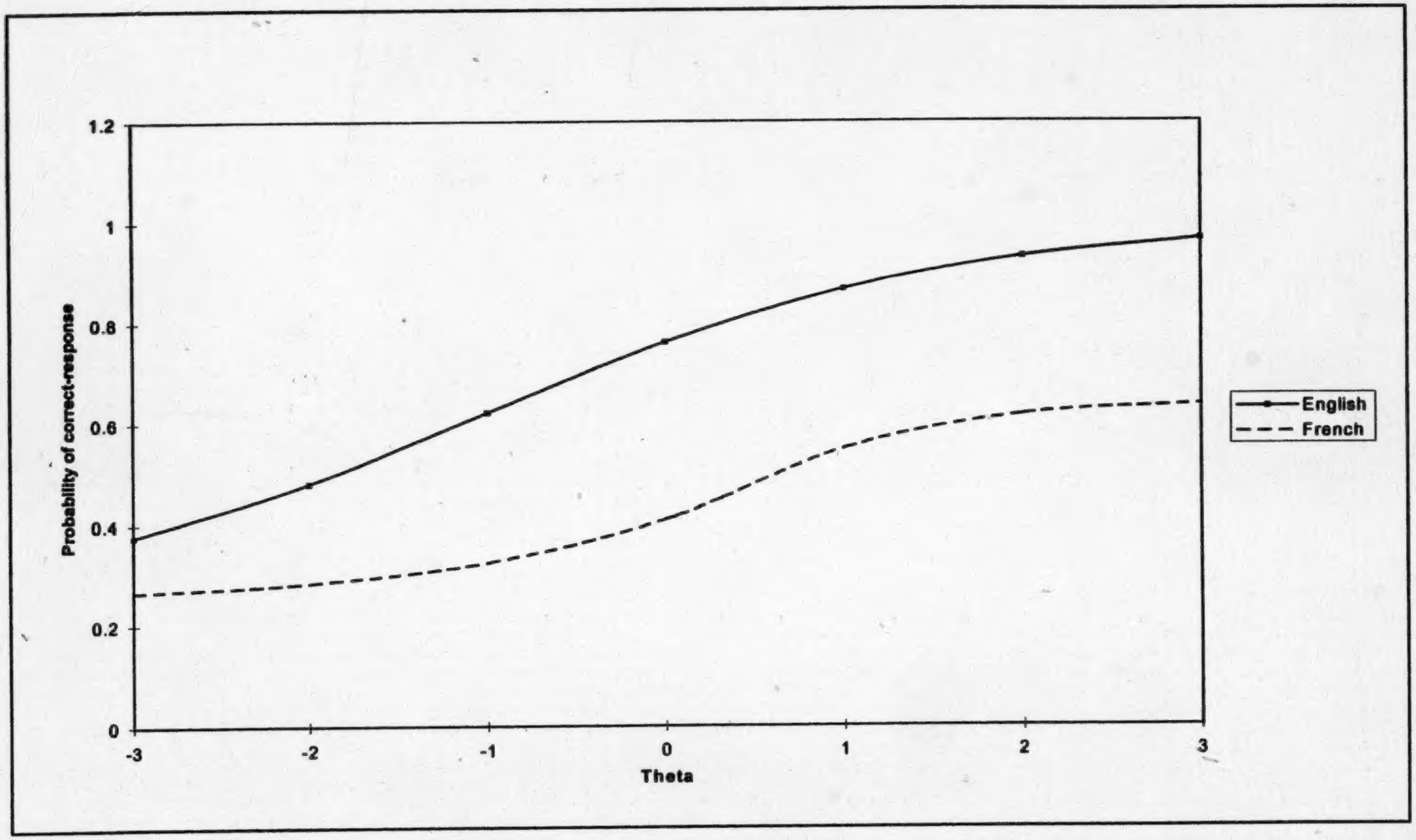

Probability of correct-response
1.2
1
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
Theta
English
French

Figure 2.

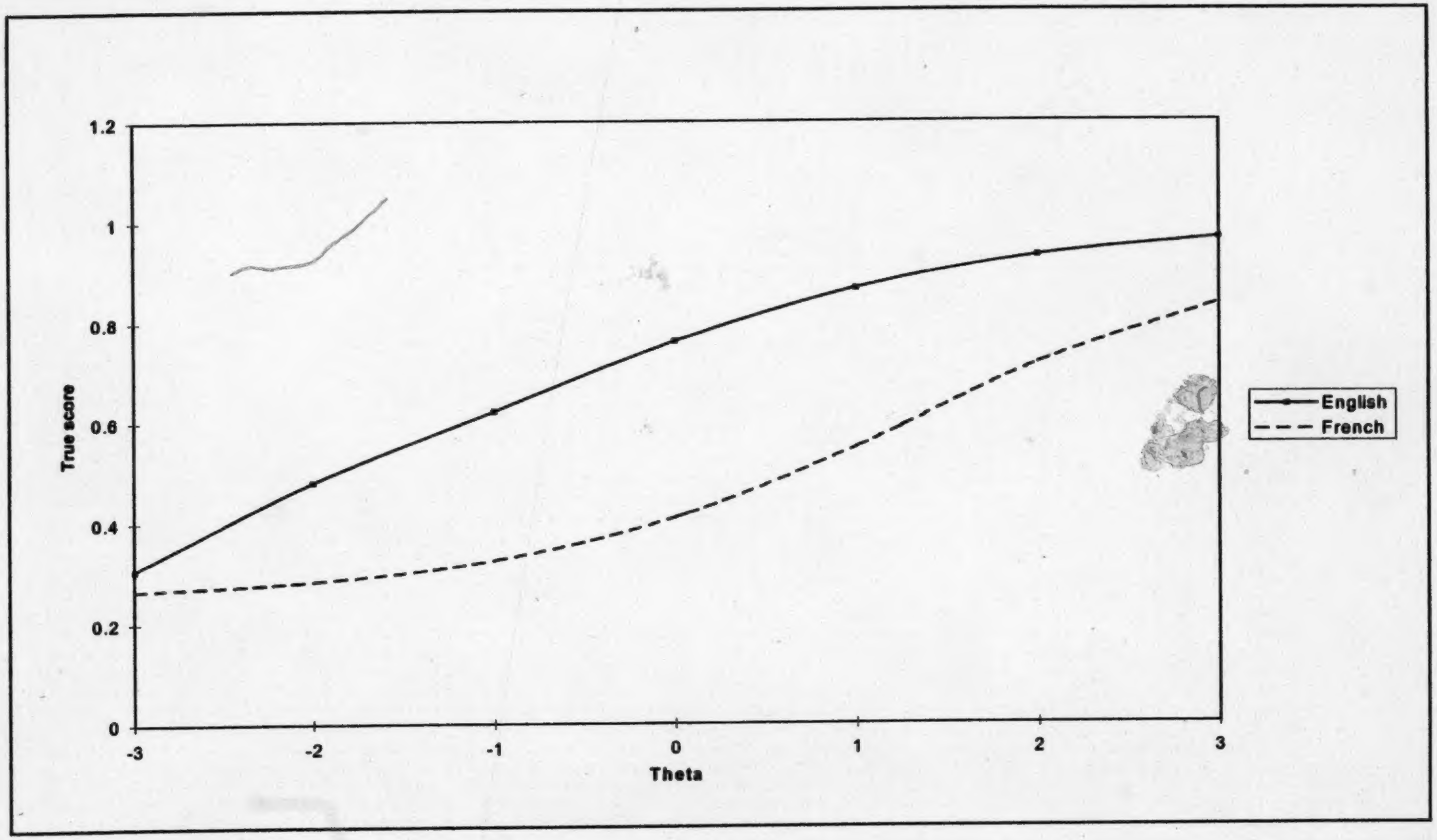


Figure 3.

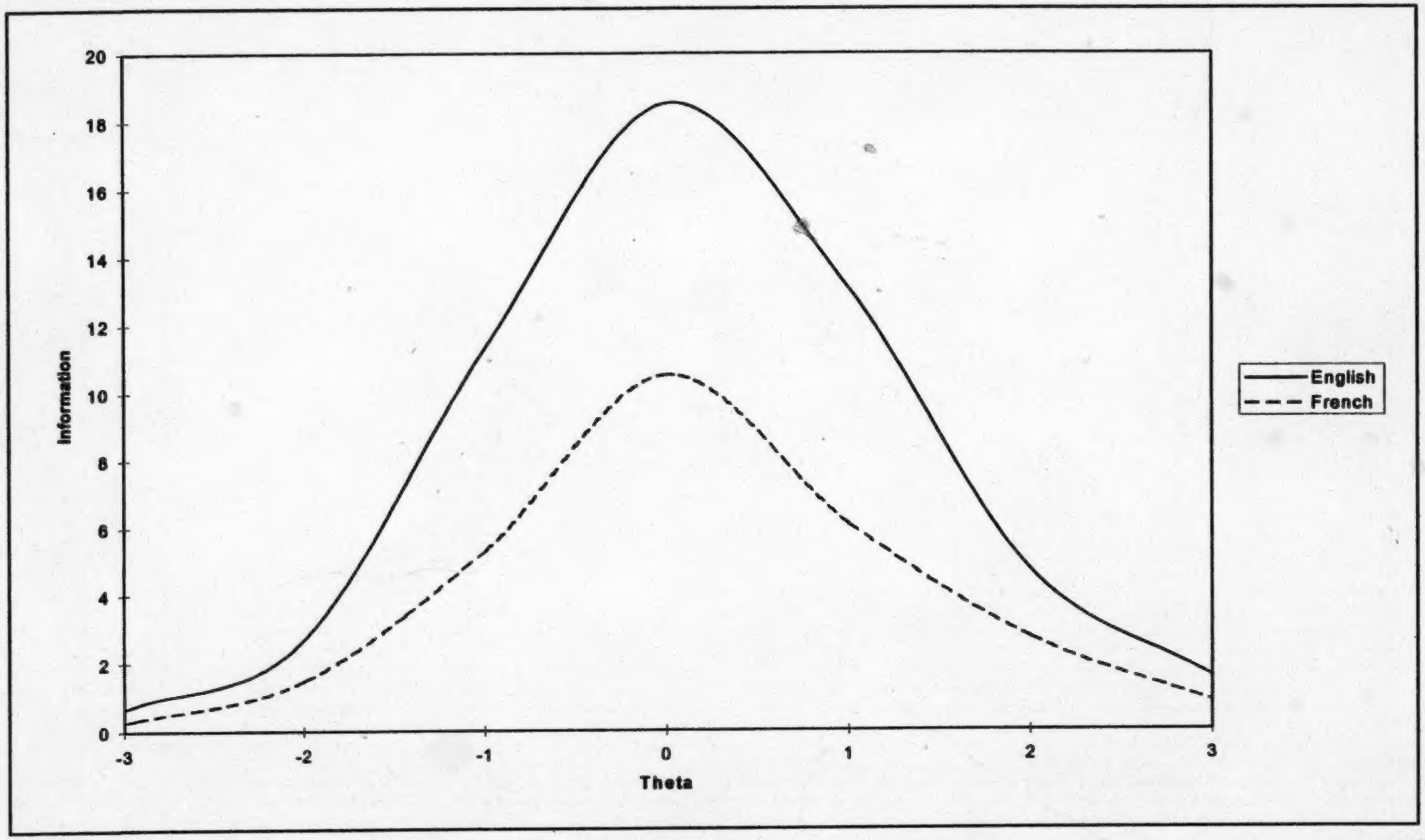

20
18
16
14
12
10
8
6
4
2
0
Information
-3
-2
-1
0
1
2
3
Theta
English
French

Figure 4.

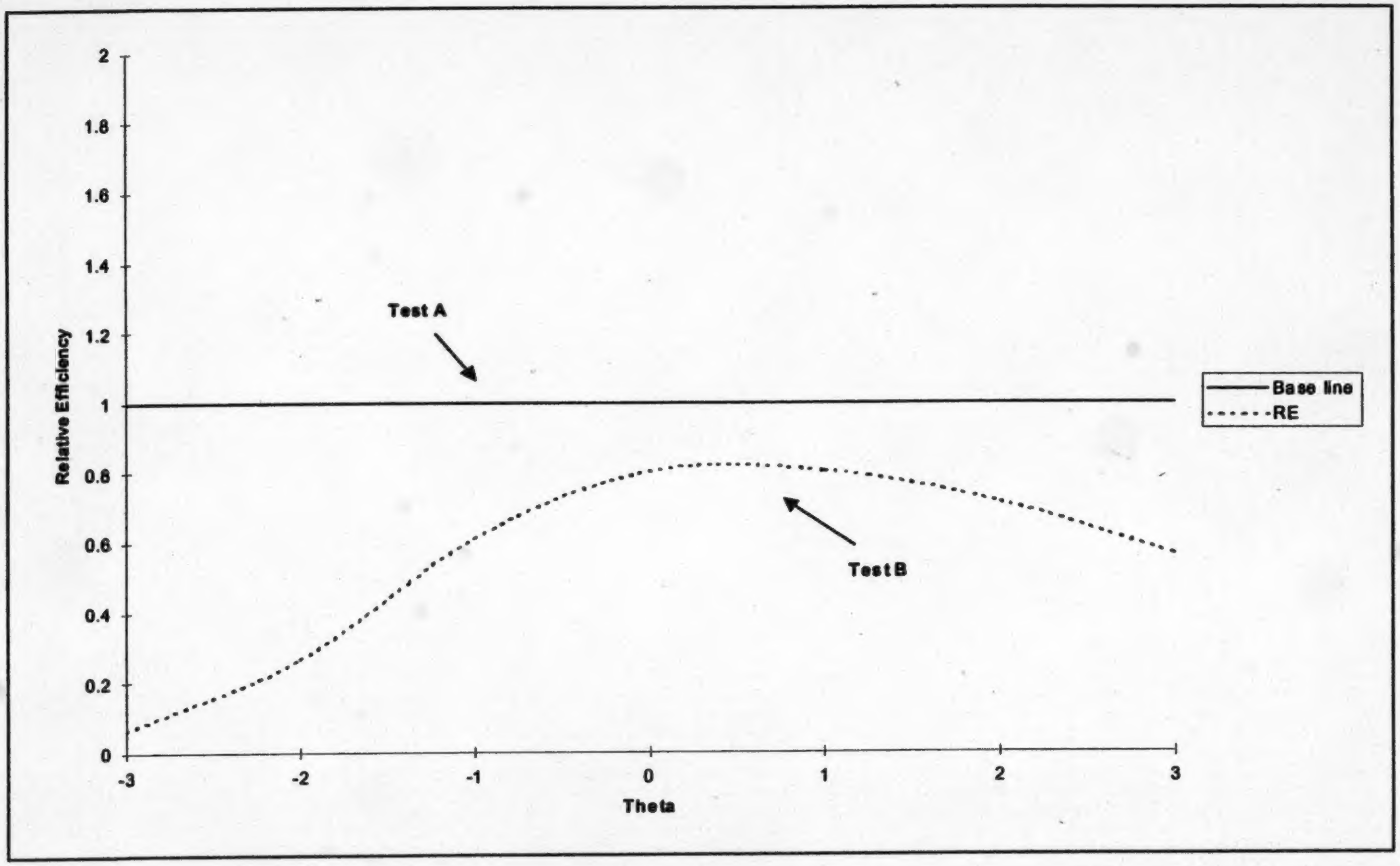

Relative Efficiency
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
Theta
Test A
Test B
Base line
RE

Figure 5.

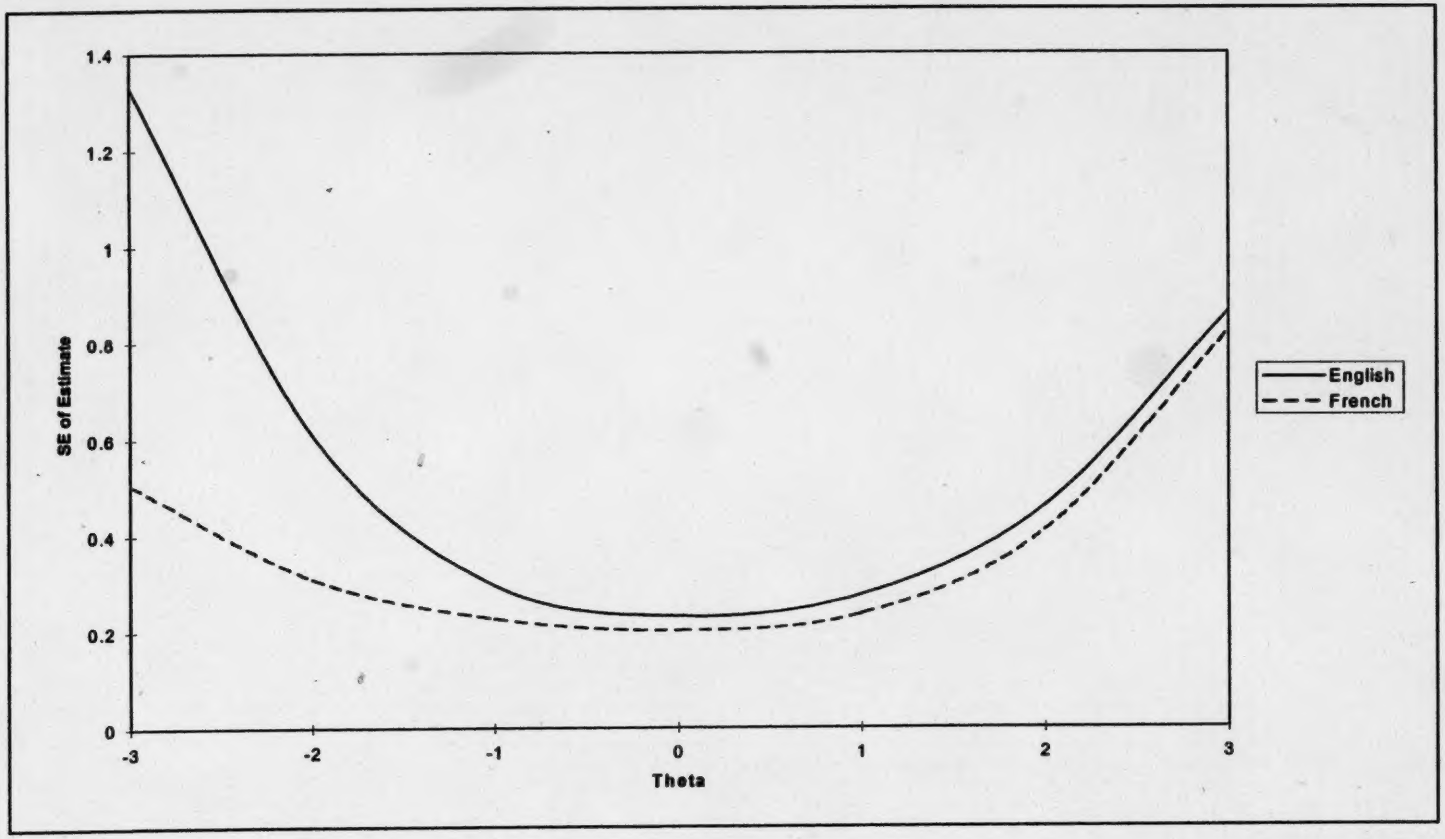
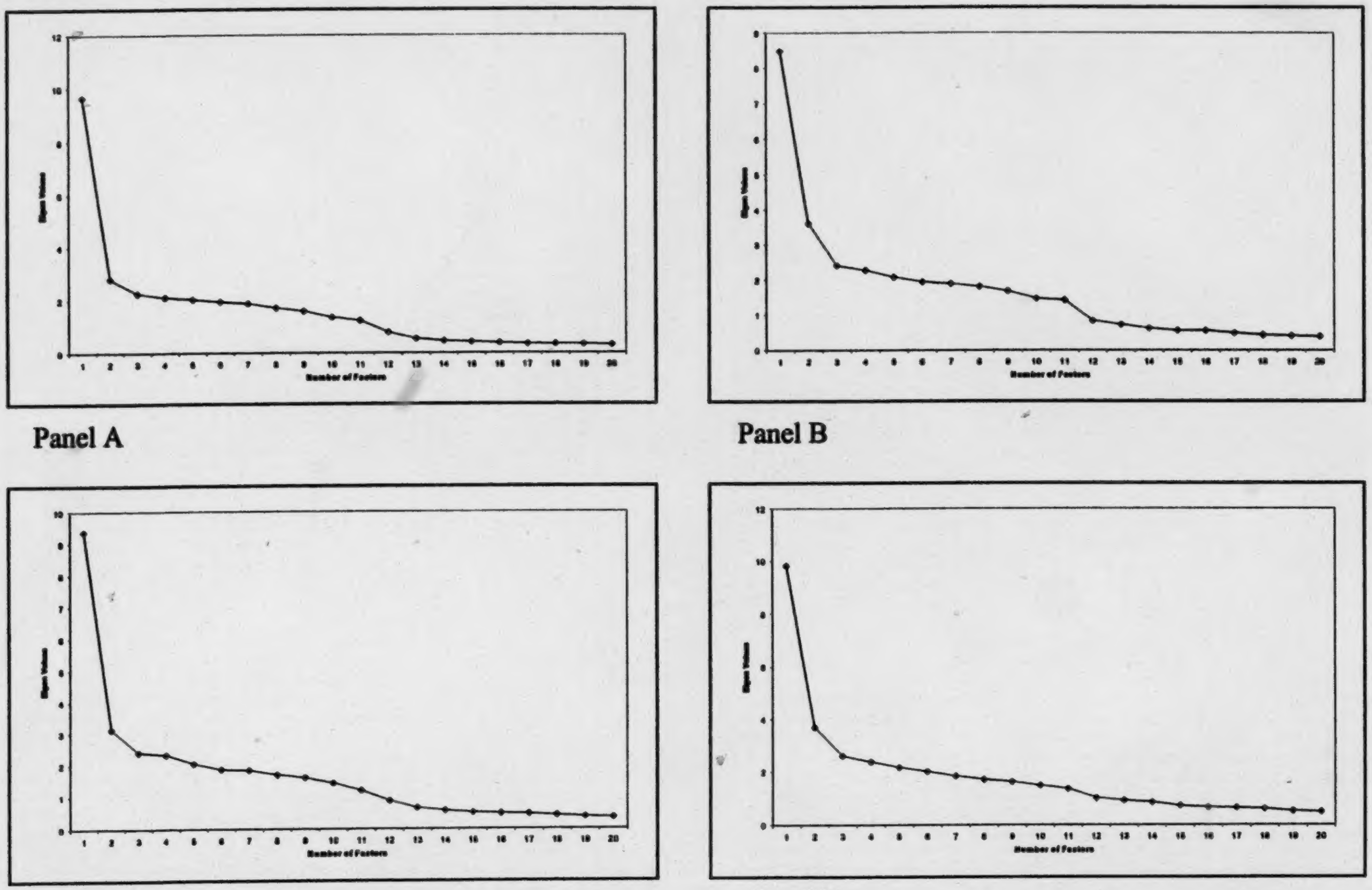

SE of Estimate
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
Theta
English
French

Figure 6.



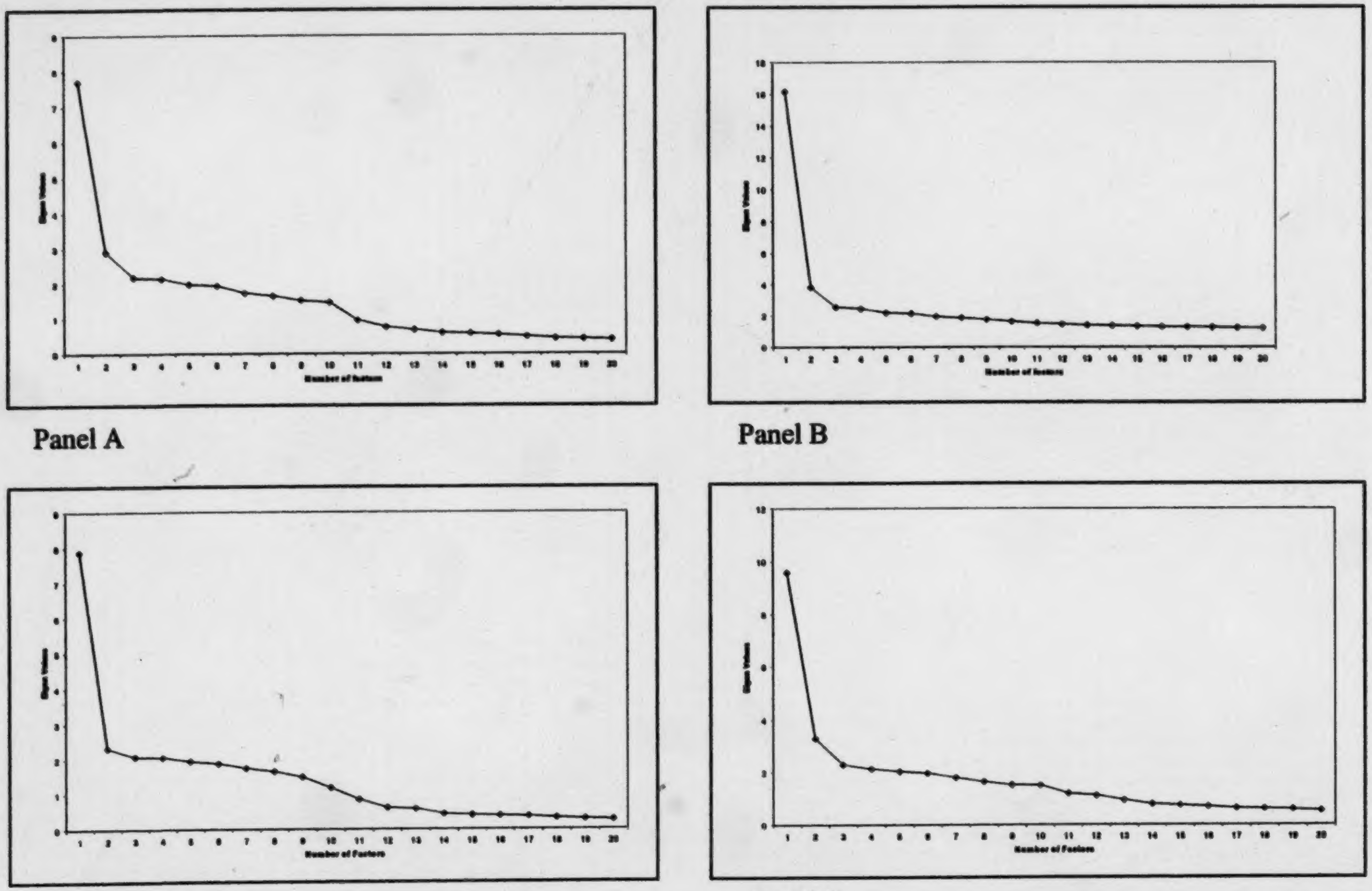
Panel A

Panel B

Panel C

Panel D

Figure 7.



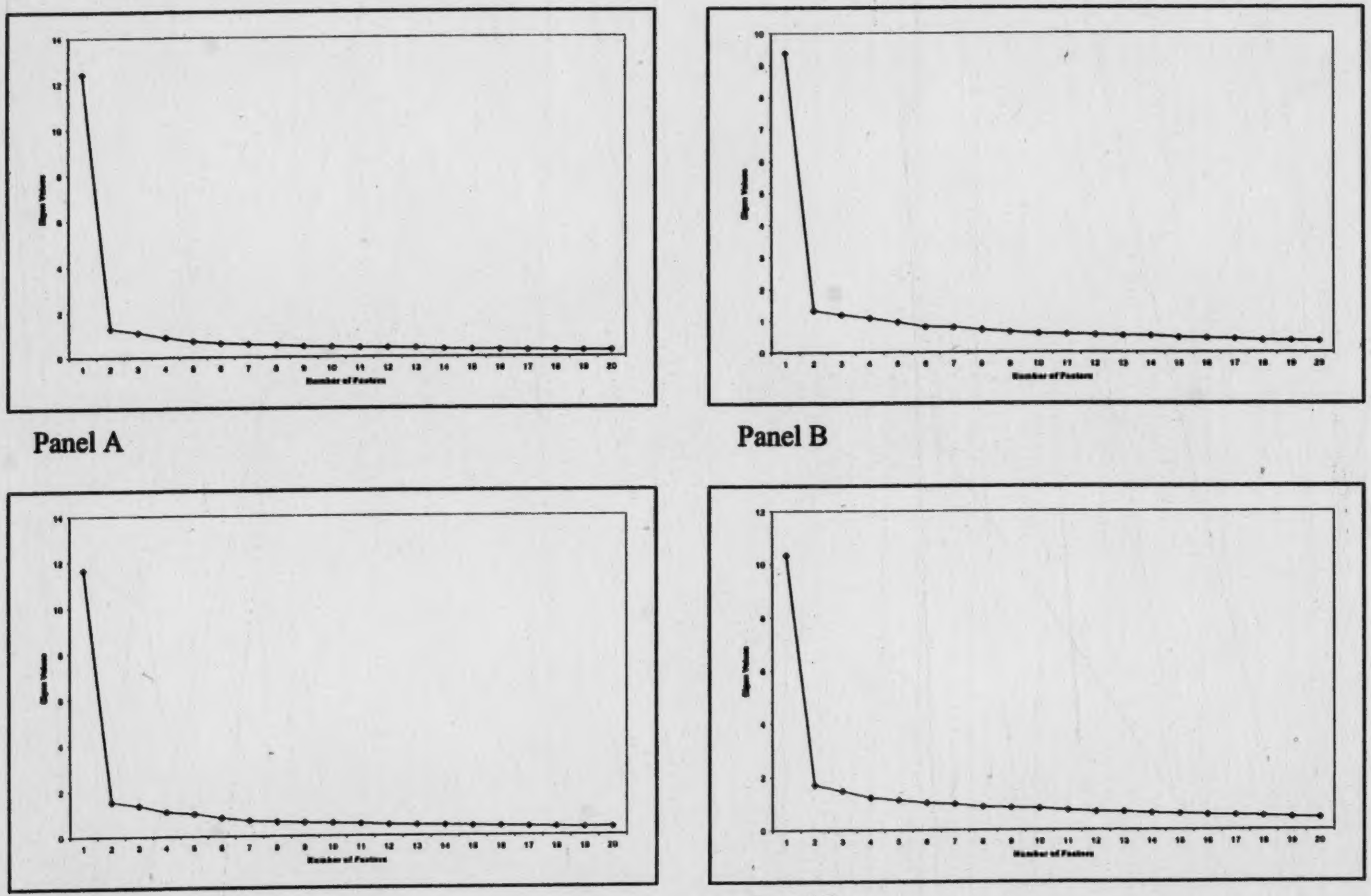
Panel A

Panel B

Panel C

Panel D

Figure 8.



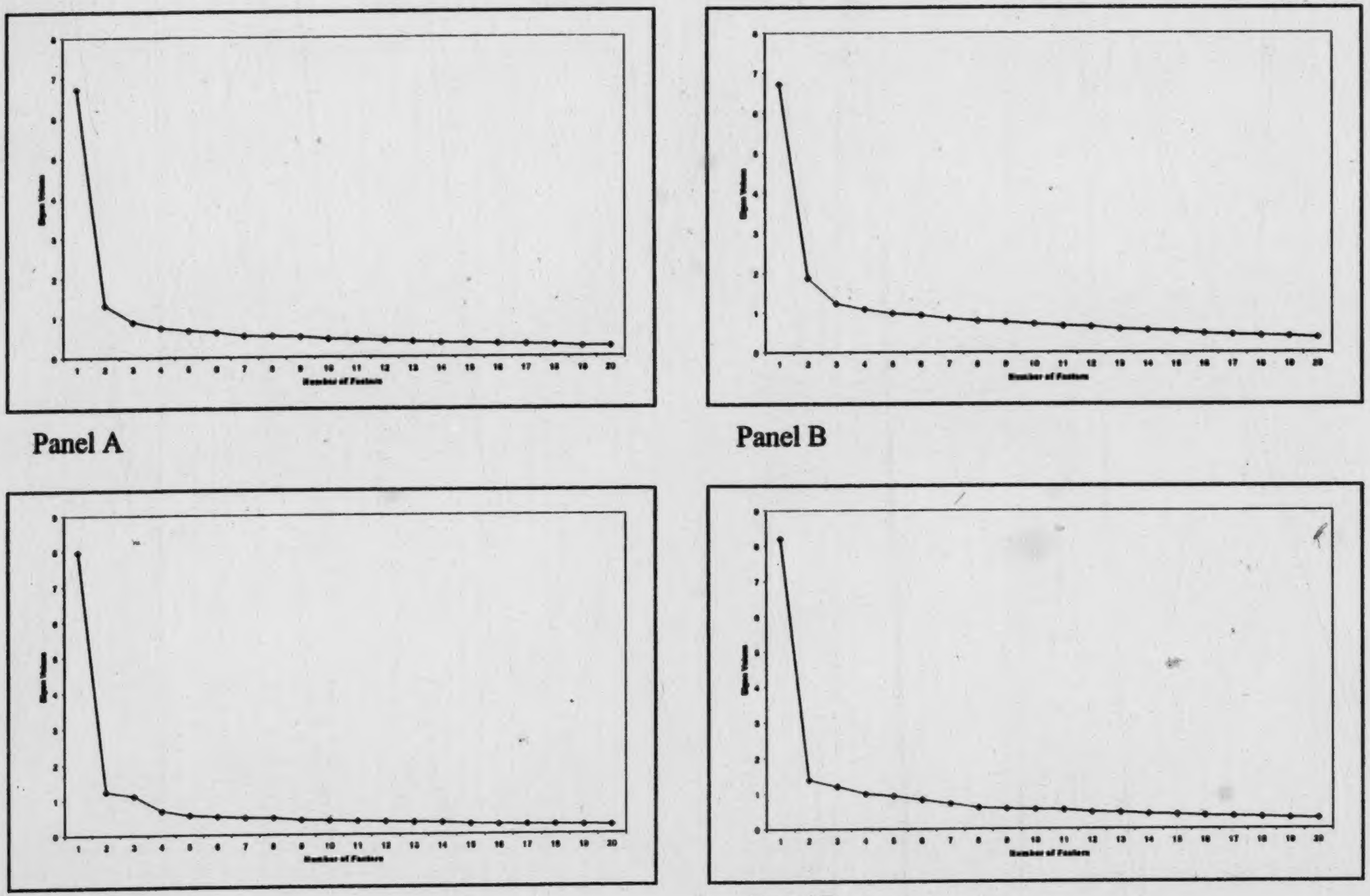
Panel A

Panel B

Panel C

Panel D

Figure 9.



Panel A

Panel B

Panel C

Panel D

Figure 10.

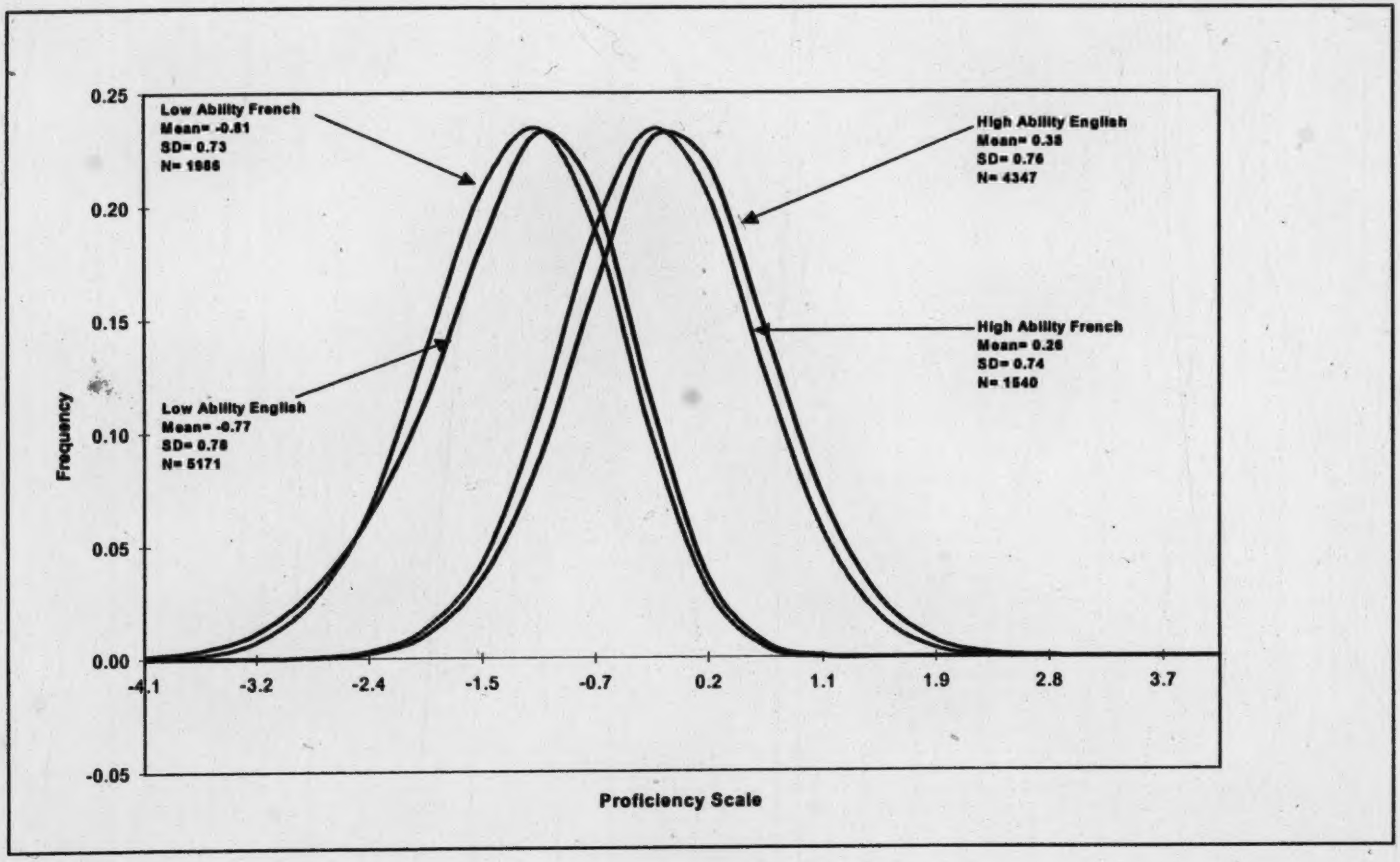

Low Ability French
Mean= -0.81
SD= 0.73
N= 1986
Low Ability English
Mean= -0.77
SD= 0.78
N= 5171
High Ability English
Mean= 0.38
SD= 0.76
N= 4347
High Ability French
Mean= 0.26
SD= 0.74
N= 1540
Frequency
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-4.1
-3.2
-2.4
-1.5
-0.7
0.2
1.1
1.9
2.8
3.7
Proficiency Scale

Figure 11.

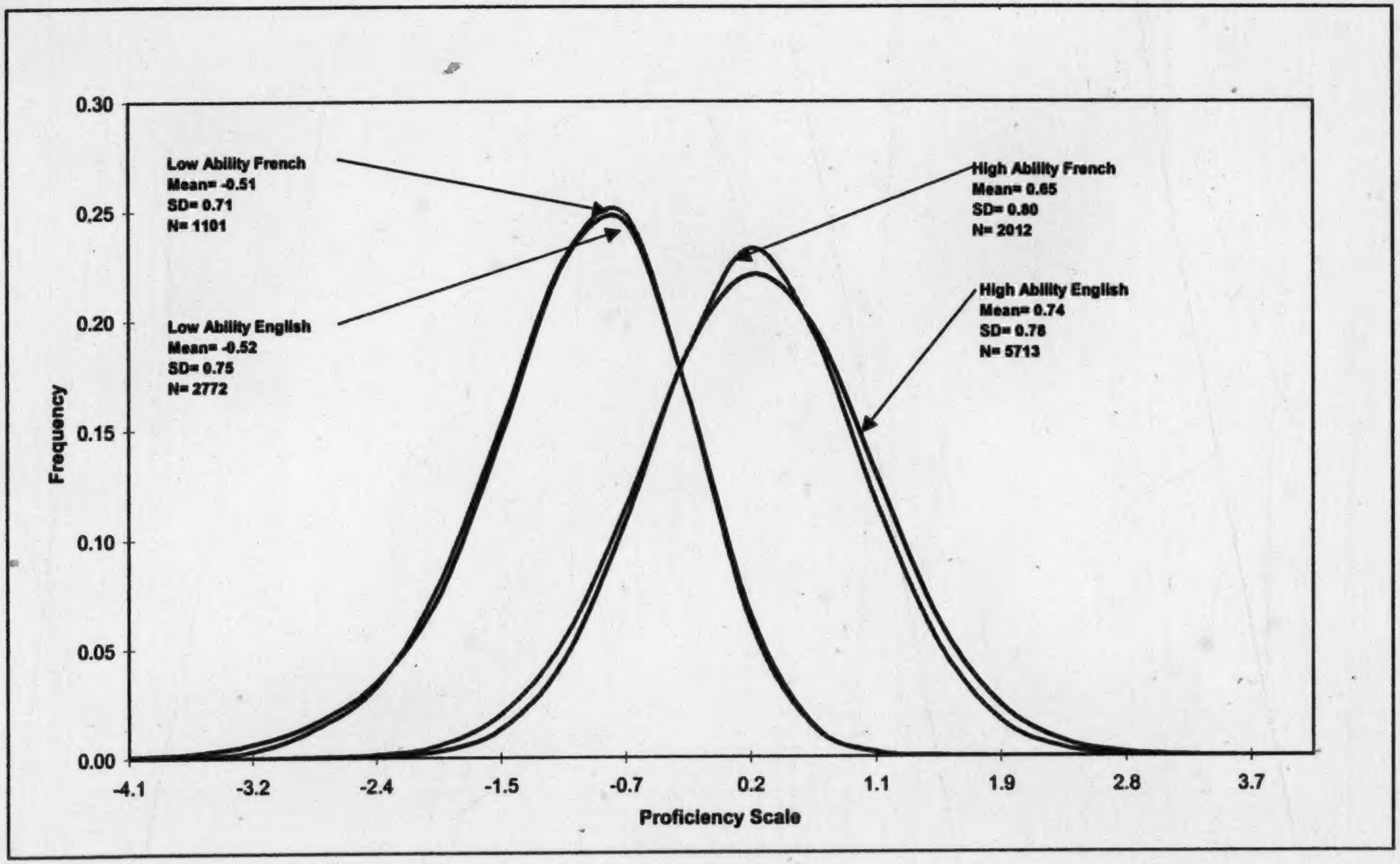

Frequency
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-4.1
-3.2
-2.4
-1.5
-0.7
0.2
1.1
1.9
2.8
3.7
Proficiency Scale
Low Ability French
Mean= -0.51
SD= 0.71
N= 1101
Low Ability English
Mean= -0.52
SD= 0.75
N= 2772
High Ability French
Mean= 0.65
SD= 0.80
N= 2012
High Ability English
Mean= 0.74
SD= 0.78
N= 5713

Figure 12.

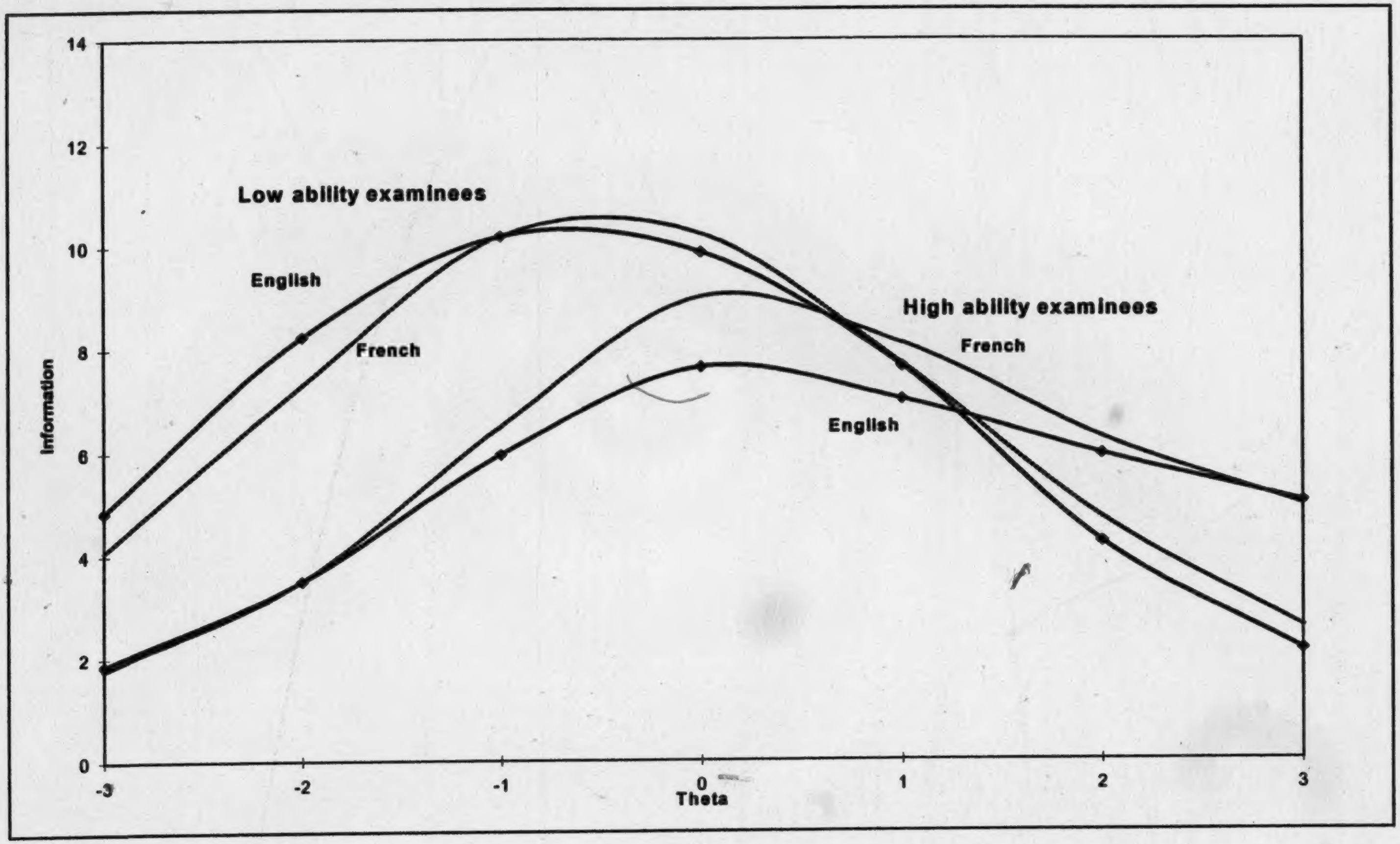

Information
14
12
10
8
6
4
2
0
-3
-2
-1
0
1
2
3
Theta
Low ability examinees
English
French
High ability examinees
French
English

Figure 13.

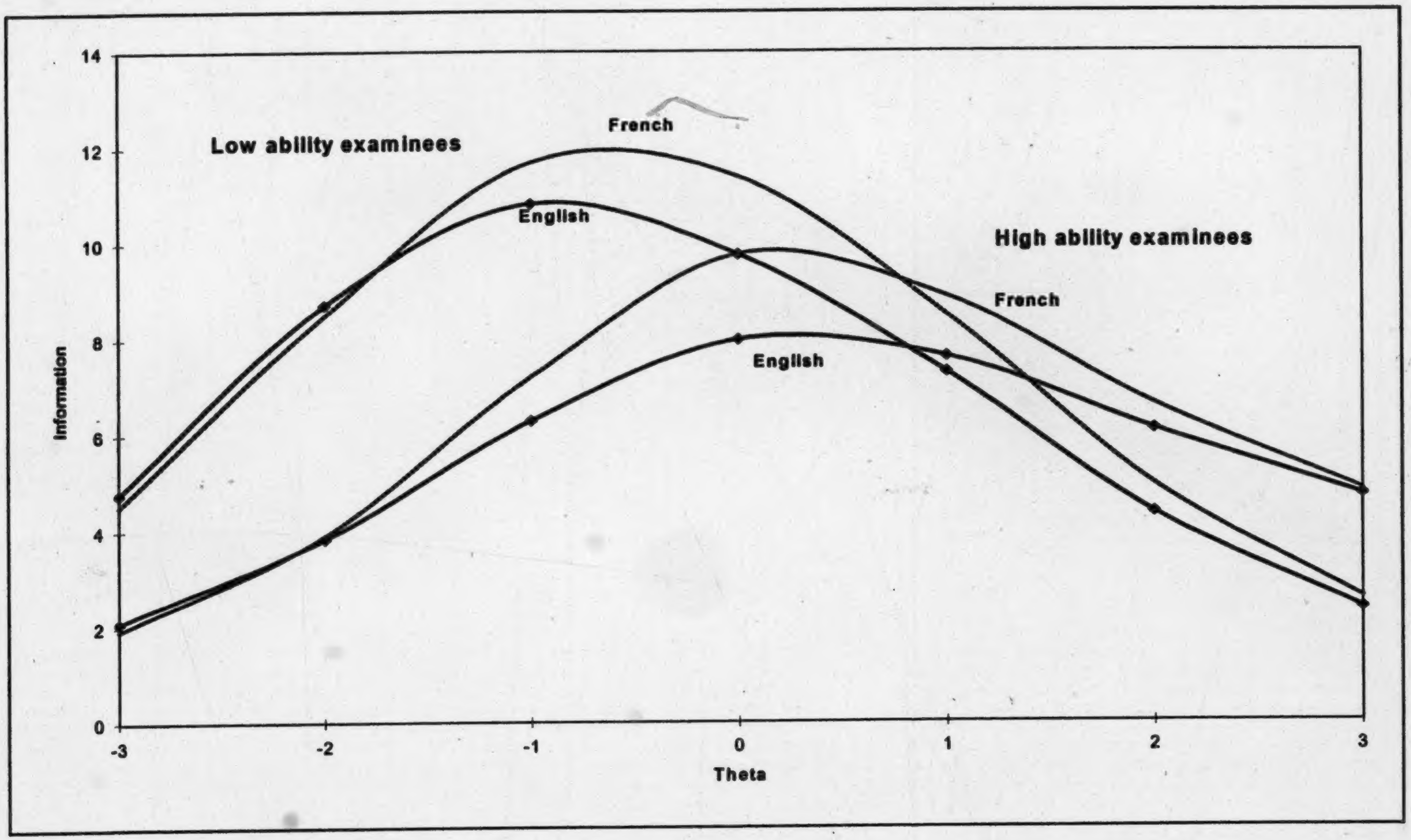
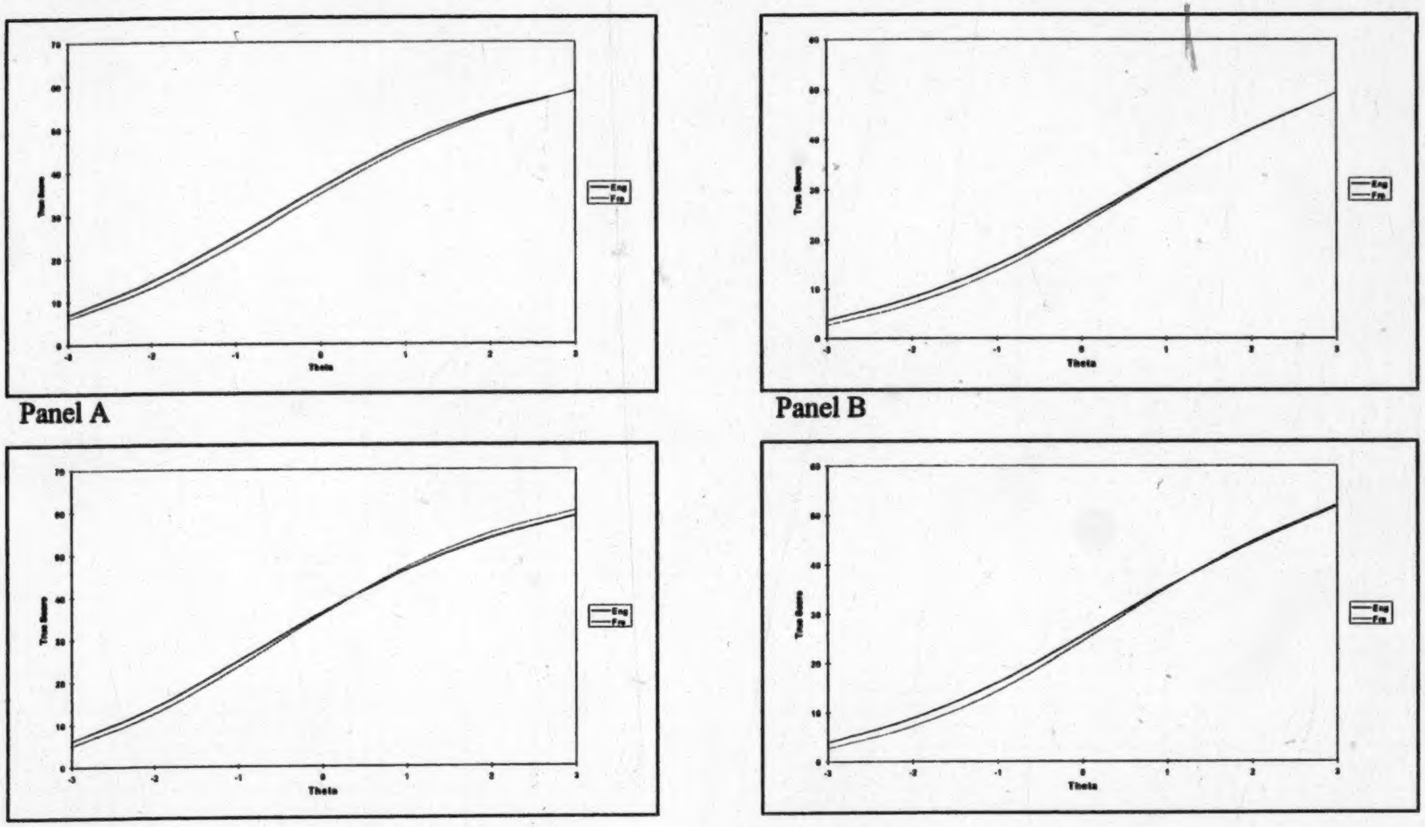

Low ability examinees
French
English
High ability examinees
French
English
Information
14
12
10
8
6
4
2
0
-3
-2
-1
0
1
2
3
Theta

Figure 14.



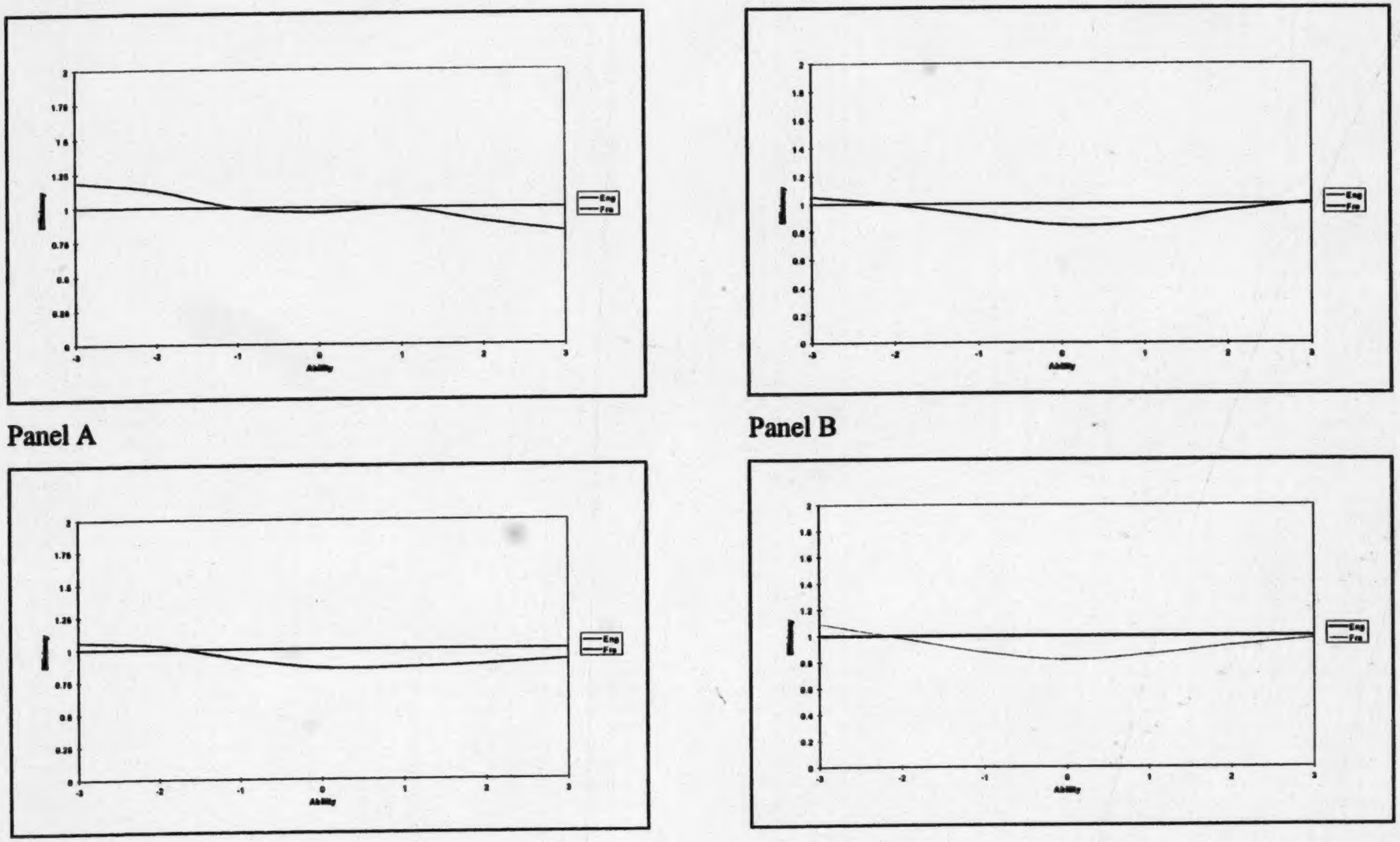
Panel A

Panel B

Panel C

Panel D

Figure 15.



Panel A

Panel B

Panel C

Panel D

Figure 16.

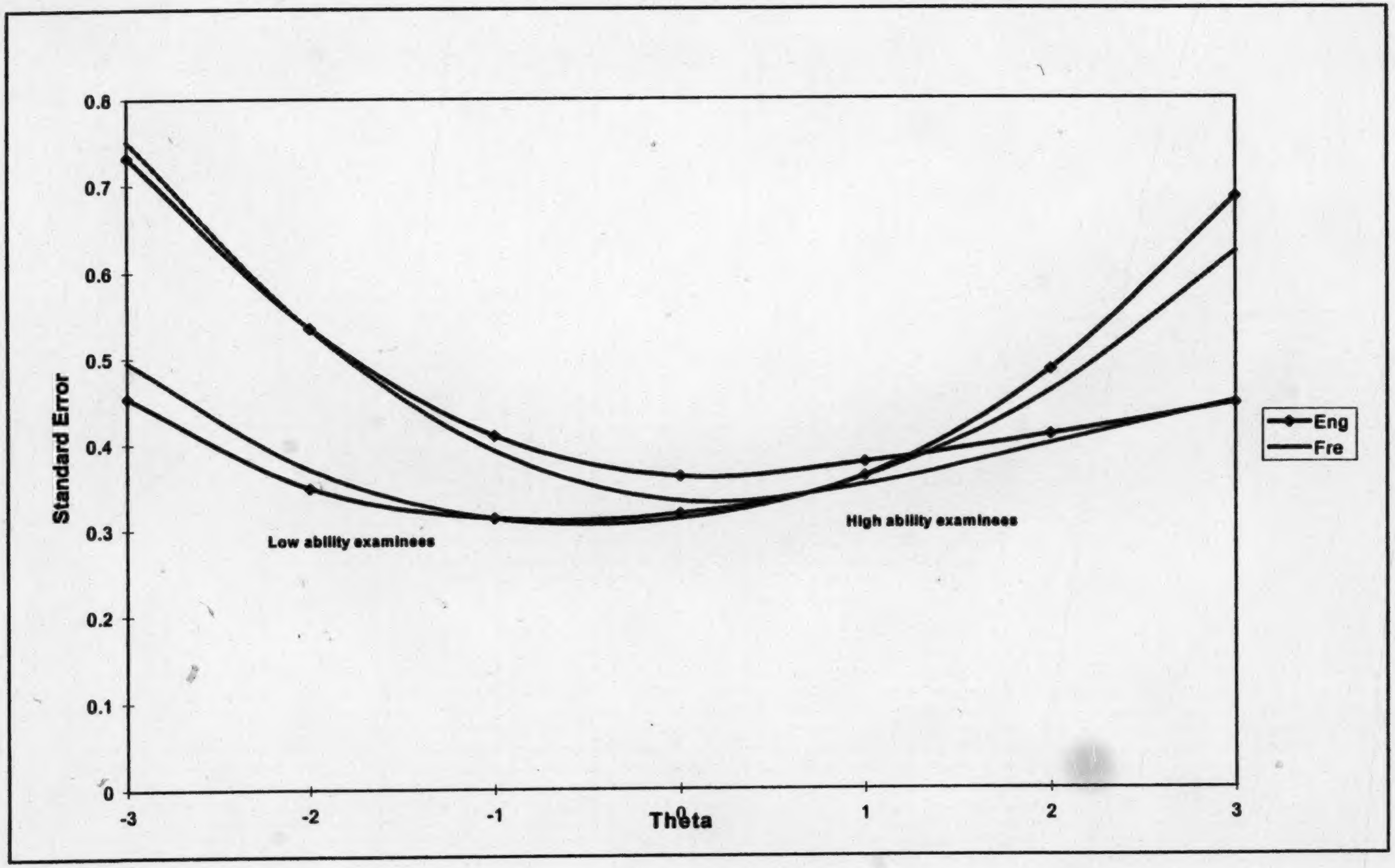

Standard Error
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-3
-2
-1
0
1
2
3
Theta
Low ability examinees
High ability examinees
Eng
Fre

Figure 17.

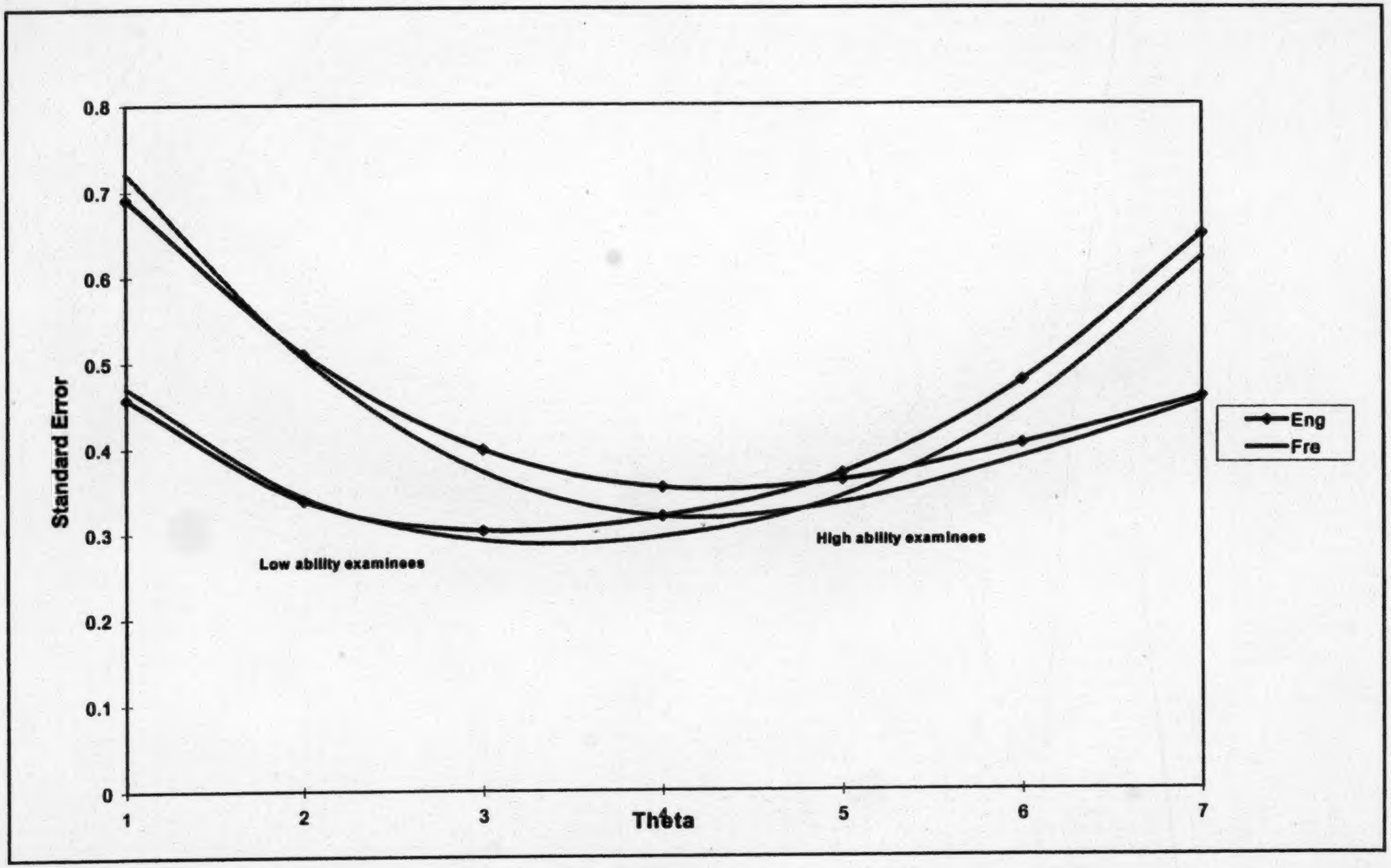

Standard Error
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
Theta
Low ability examinees
High ability examinees
Eng
Fre

Figure 18.

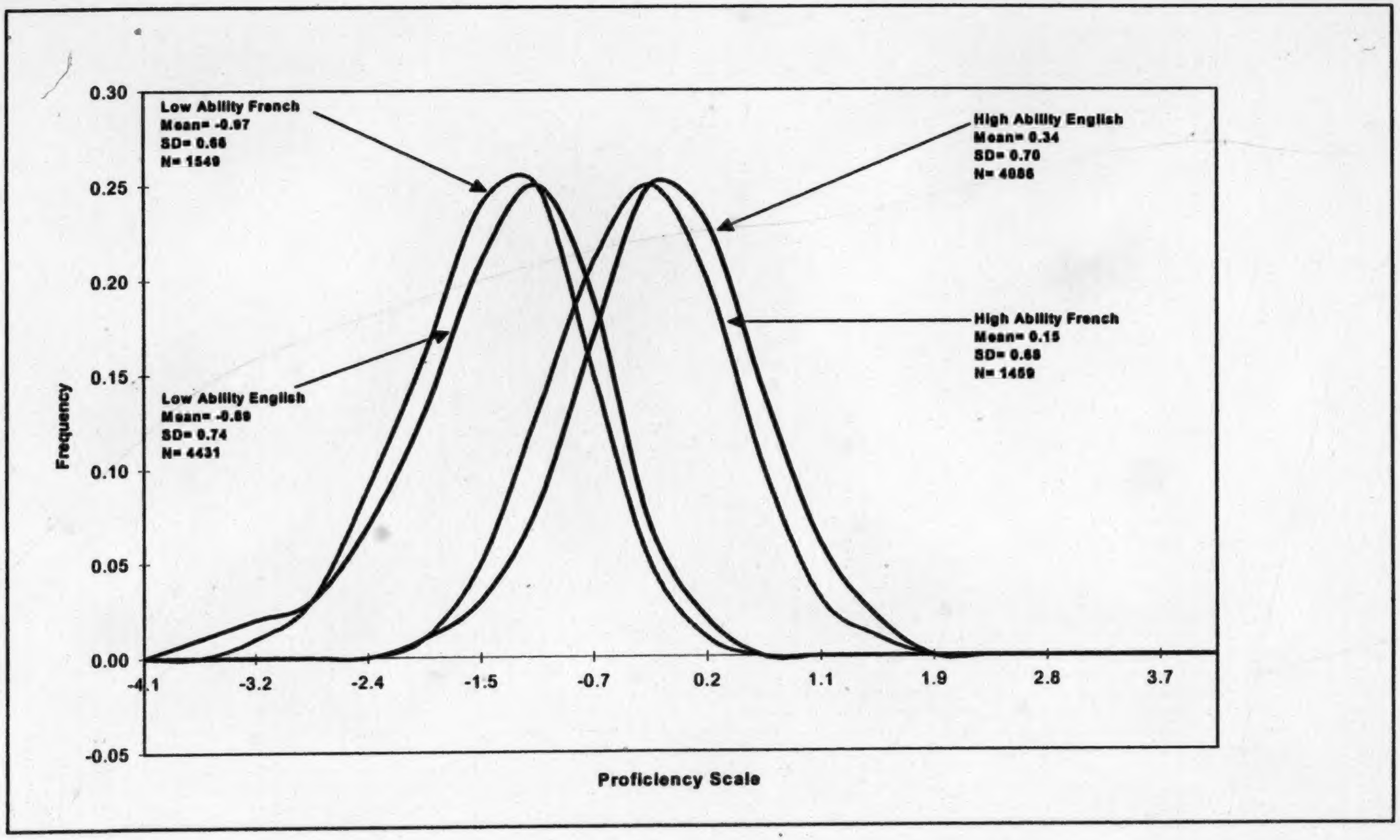

0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
Frequency
-4.1
-3.2
-2.4
-1.5
-0.7
0.2
1.1
1.9
2.8
3.7
Proficiency Scale
Low Ability French
Mean= -0.97
SD= 0.66
N= 1549
Low Ability English
Mean= -0.89
SD= 0.74
N= 4431
High Ability English
Mean= 0.34
SD= 0.70
N= 4086
High Ability French
Mean= 0.15
SD= 0.68
N= 1459

Figure 19.

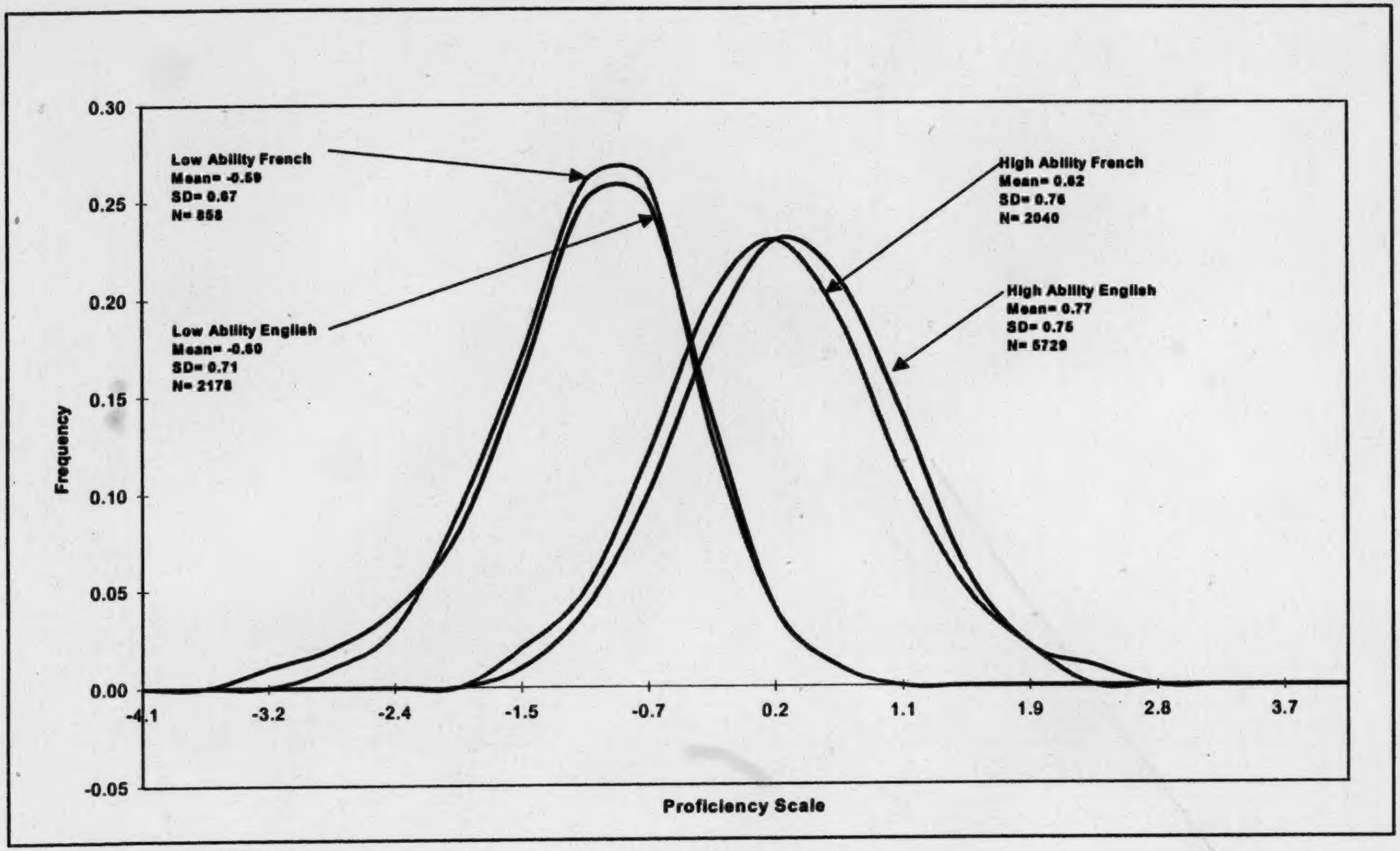

Low Ability French
Mean= -0.59
SD= 0.67
N= 858
Low Ability English
Mean= -0.60
SD= 0.71
N= 2178
High Ability French
Mean= 0.62
SD= 0.76
N= 2040
High Ability English
Mean= 0.77
SD= 0.75
N= 5729
Frequency
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-4.1
-3.2
-2.4
-1.5
-0.7
0.2
1.1
1.9
2.8
3.7
Proficiency Scale

Figure 20.

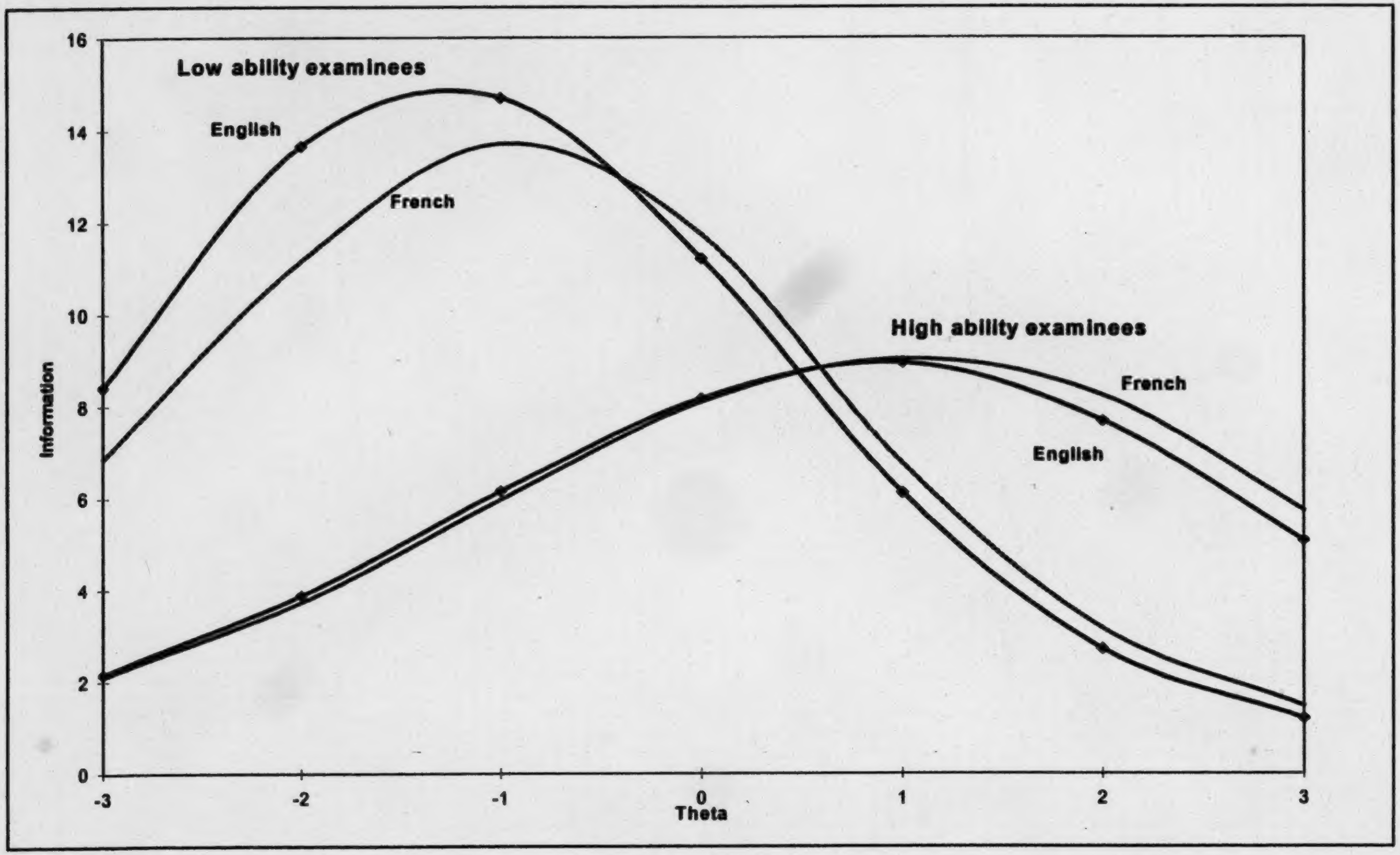

Low ability examinees
English
French
High ability examinees
French
English
Information
16
14
12
10
8
6
4
2
0
-3
-2
-1
0
1
2
3
Theta

Figure 21.

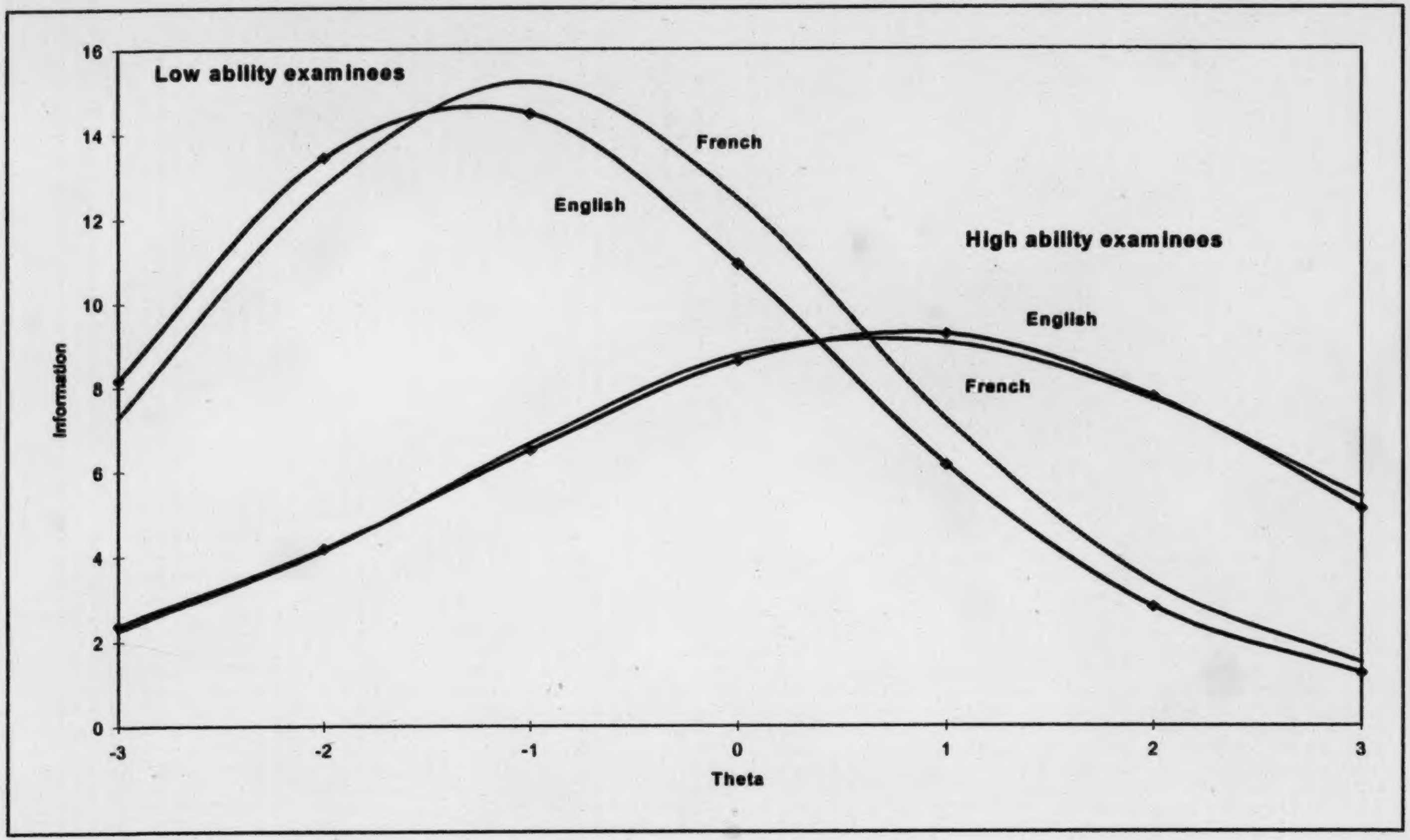
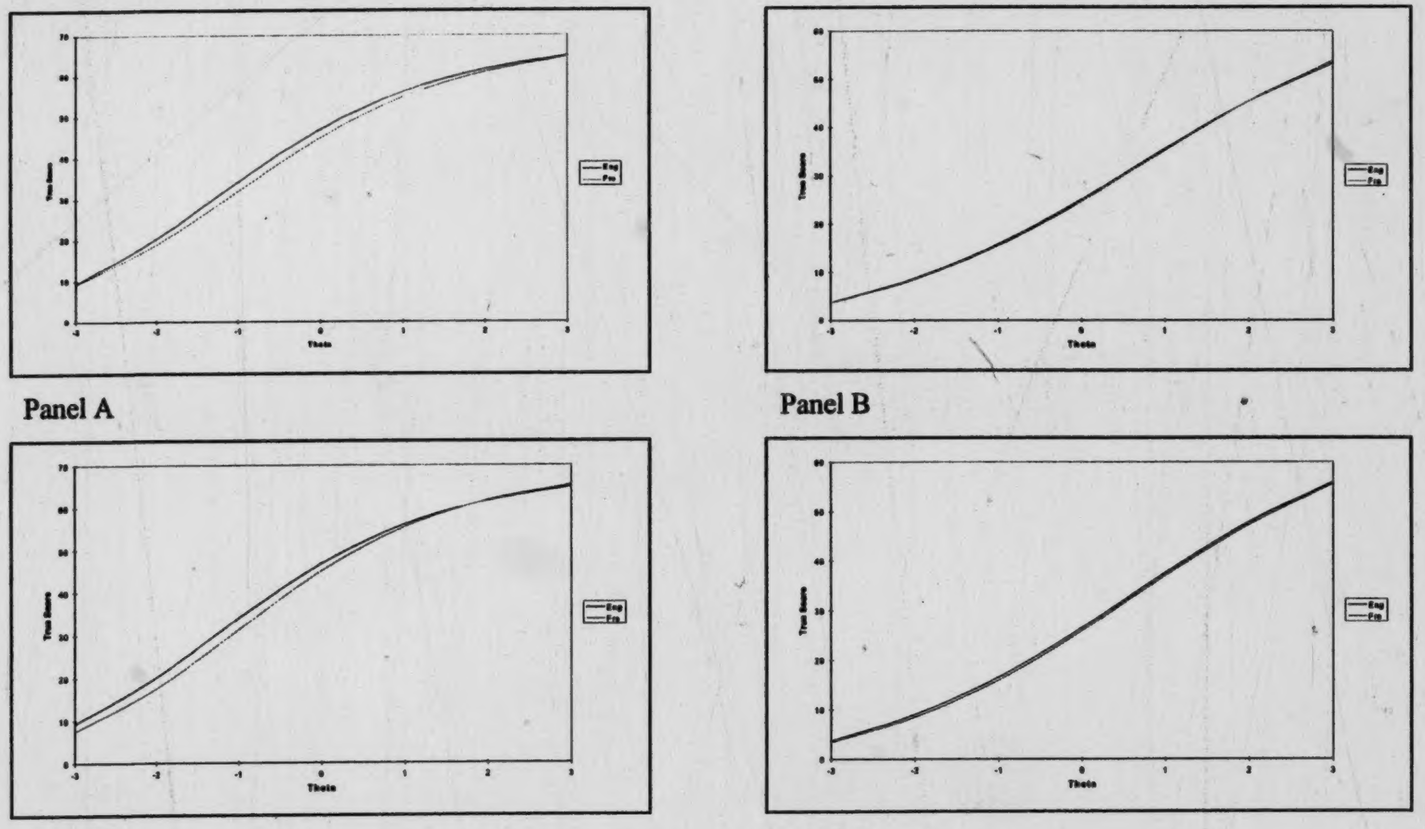

Low ability examinees
French
English
High ability examinees
English
French
Information
16
14
12
10
8
6
4
2
0
-3
-2
-1
0
1
2
3
Theta

Figure 22.



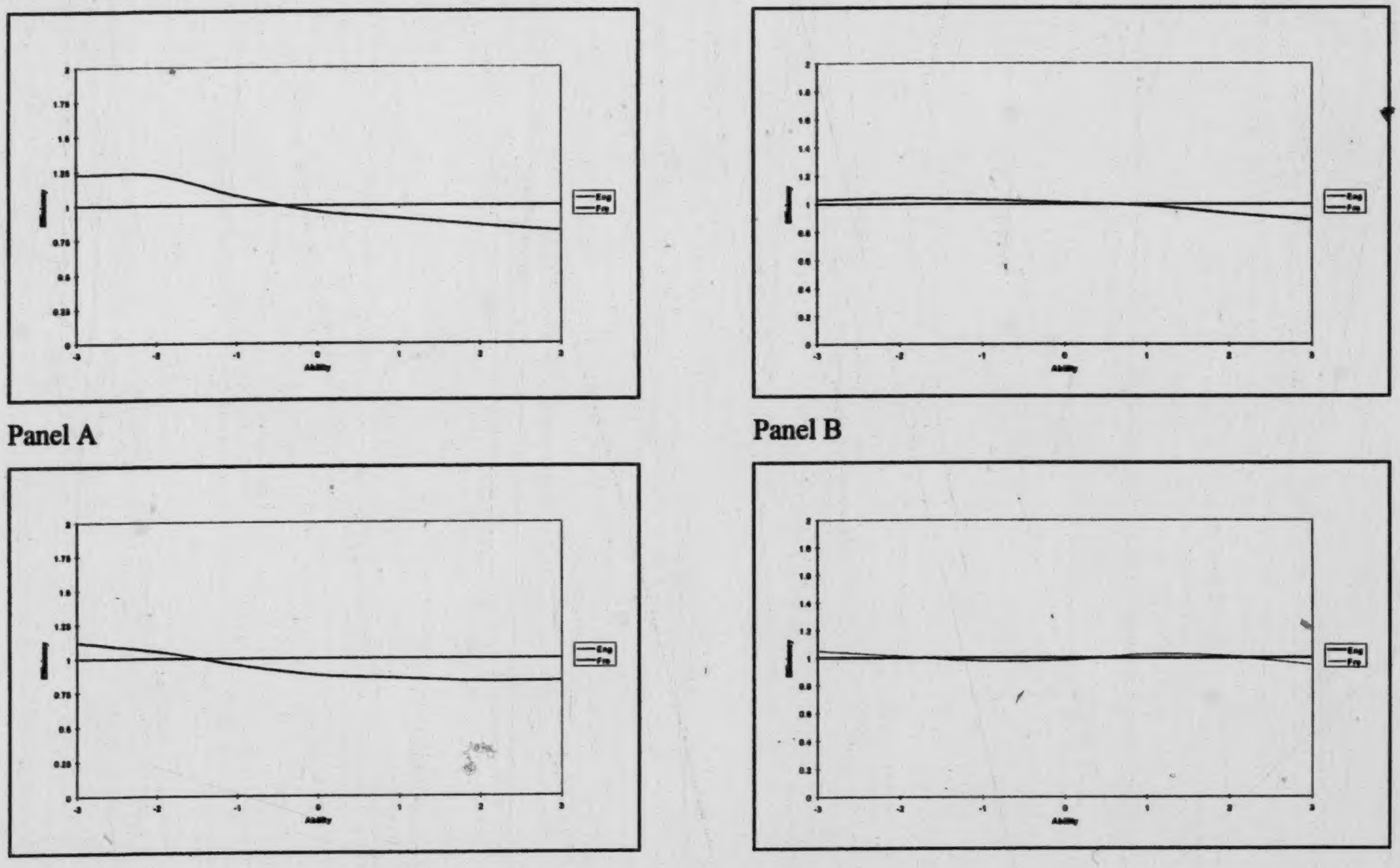
Panel A

Panel B

Panel C

Panel D

Figure 23.



Panel A

Panel B

Panel C

Panel D

Figure 24.

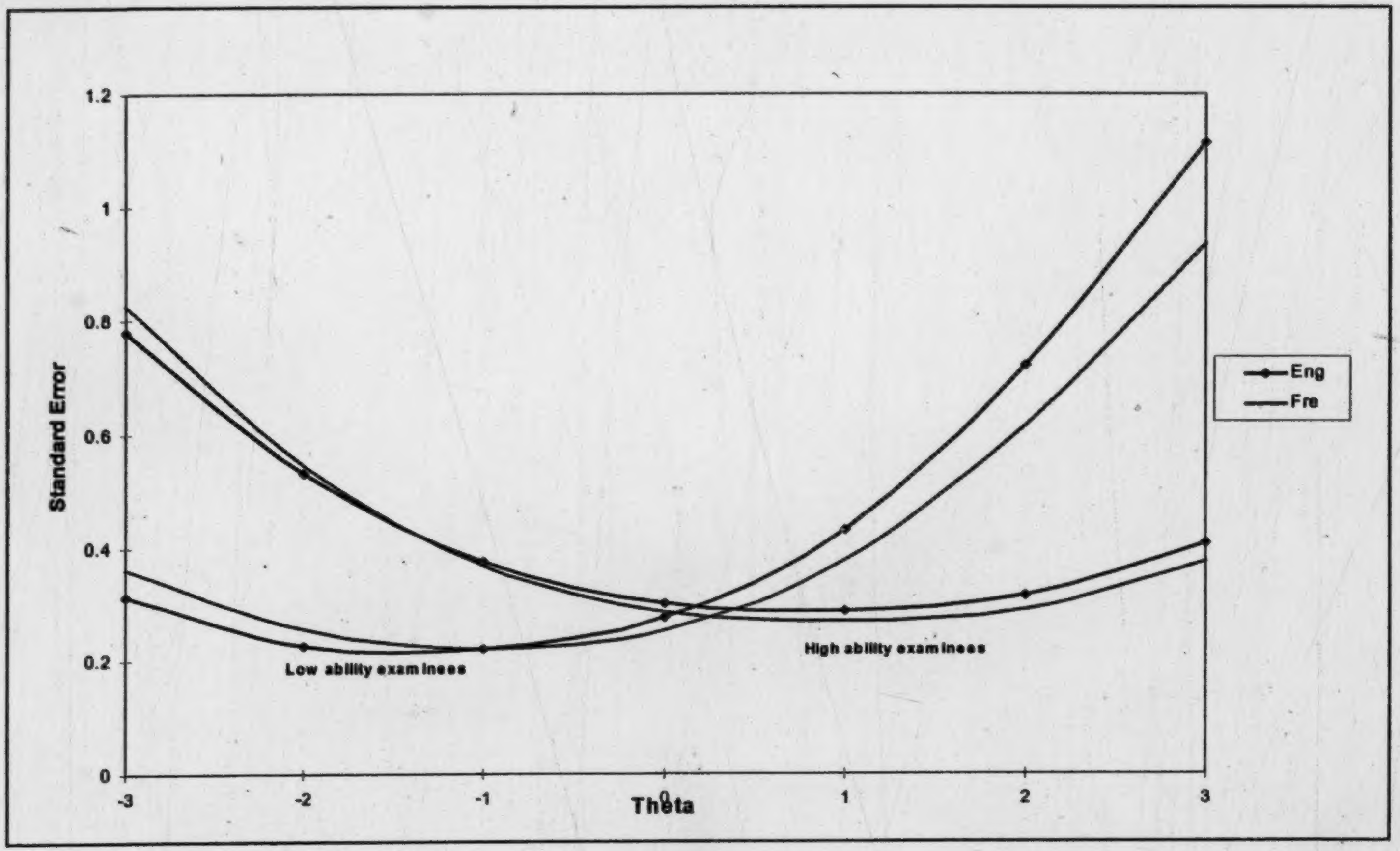
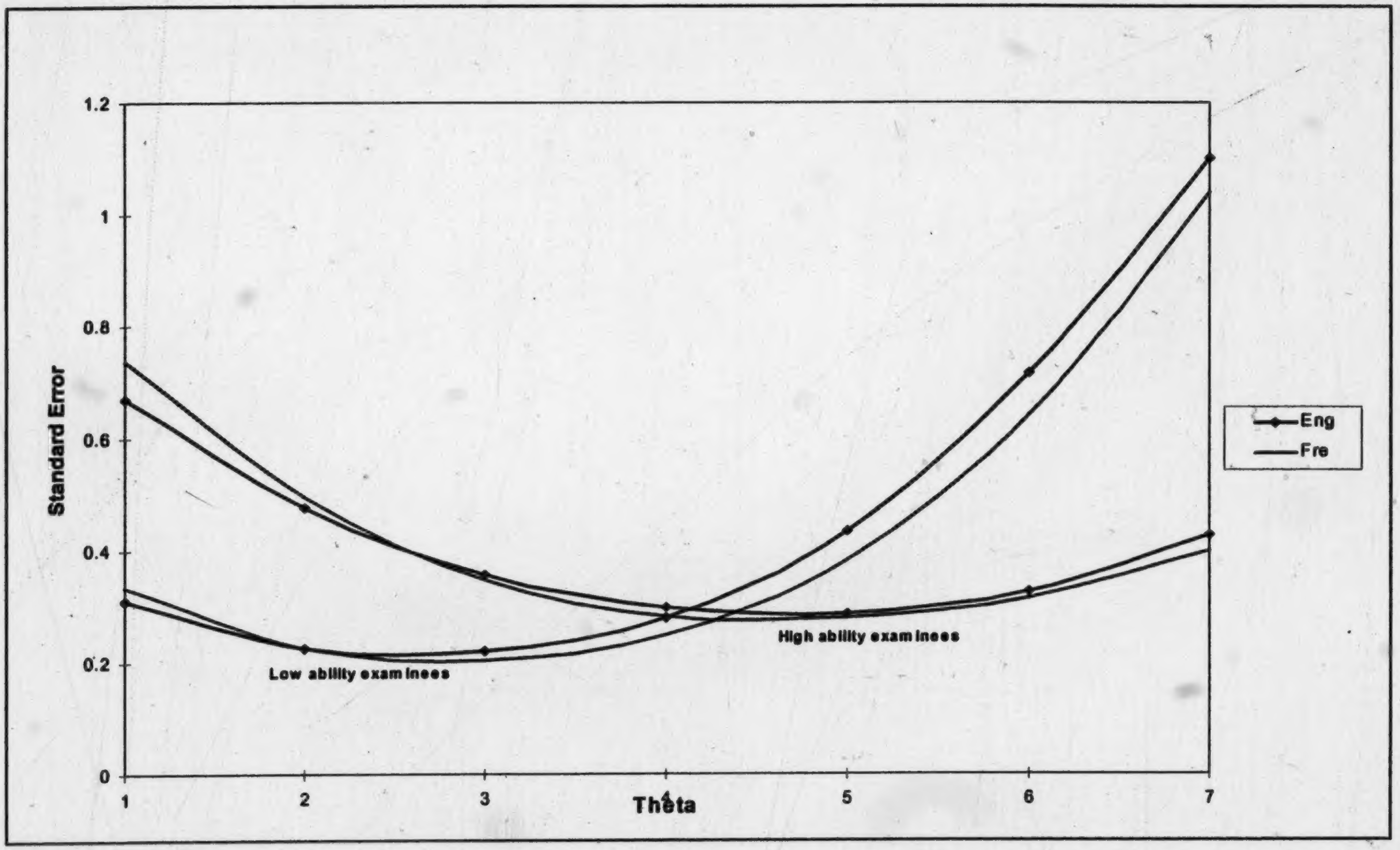

Standard Error
1.2
1
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
Theta
Eng
Fre
Low ability examinees
High ability examinees

Figure 25.


Standard Error
1.2
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
7
Thêta
Eng
Fre
Low ability examinees
High ability examinees

Appendix A

In this appendix, various achievement levels and sample questions from the SAIP Science tests are described. *At level one*, students are expected to complete tasks such as describing physical properties of objects, distinguishing living things from non-living things, and identifying various technologies important to society.

Sample level one question:

While on the trip, students will be experiencing more hours of sunlight than any other time of the year. During which month are they going?

A. March

*B. June**

C. September

D. December

At level two, students are expected to be able to classify substances according to their physical properties, compare various plant and animal adaptations, and identify technologies that influence science, and science knowledge that leads to new technologies.

Sample level two question:

Which of the following describes a behavior that birds would have learned?

A. Bringing food back to the nest to feed their young

*B. Looking for food near campsites**

C. Sleeping while perched on a branch

D. Building nests with small twigs

At level three, students are expected to complete tasks such as using chemical properties to compare and classify substances, analyzing experiments and judging their validity, and identifying areas where science knowledge and technologies address societal problems.

Sample level three question:

Michelle knows that the light from the Moon's surface reaches Earth in about one second. She knows that the light from the Alpha Centauri, the star nearest our solar system takes five years to reach Earth. About how long does it take for light to travel from Sun to Earth?

A. 1 second

*B. 8 minutes**

C. 5 years

D. 10 years

At level four, students are expected to complete tasks such as describing and comparing particles in terms of protons, neutrons, and electrons and explaining that scientific progress is the result of ongoing experimentation and evaluation.

Sample level four question:

Oxygen is an important component of air found in soil. Like many other substances oxygen is cycled in nature. Describe the oxygen cycle in nature? Use a labeled diagram if you wish.

At *level five*, students are expected to relate properties of substances to their molecular structure, analyze uniform motion in two dimensions, and explain conditions used to evaluate scientific theories.

Sample level five question:

Methane is another fuel used in homes. Both methane and propane are gases at room temperature and pressure. Water, on the other hand, is a liquid at room temperature and pressure. Under these conditions, why is water liquid while methane and propane are gases?

A. Methane and propane have more hydrogen making them more gaseous.

B. Methane and propane have large spaces between their molecules.

C. Water has more attractive forces between the molecules.

D. Water molecules are smaller and will pack together more tightly.

Appendix B

In this appendix, eigenvalues from the principal components analysis for the second-stage tests for the eight groups of examinees studied in the1996 administration are presented.

	Low-Ability Examinees				High-Ability Examinees			
	13-year-olds		16-year-olds		13-year-olds		16-year-olds	
	English	French	English	French	English	French	English	French
λ	9.65	8.48	9.37	9.83	7.71	16.16	7.88	9.60
	2.80	3.61	3.13	3.69	2.92	3.81	2.33	3.31
	2.26	2.41	2.40	2.60	2.19	2.57	2.08	2.30
	2.11	2.27	2.33	2.38	2.15	2.46	2.06	2.14
	2.03	2.08	2.04	2.16	1.99	2.20	1.95	2.04
	1.95	1.94	1.86	2.01	1.95	2.15	1.88	1.97
	1.88	1.89	1.83	1.84	1.74	1.96	1.75	1.80
	1.71	1.81	1.69	1.71	1.65	1.88	1.65	1.65
	1.59	1.68	1.60	1.62	1.53	1.76	1.49	1.52
	1.35	1.47	1.43	1.48	1.47	1.64	1.18	1.51
	1.22	1.43	1.20	1.36	0.95	1.52	0.85	1.20
	0.77	0.83	0.87	1.01	0.76	1.47	0.61	1.11
	0.53	0.73	0.65	0.92	0.67	1.40	0.58	0.93
	0.45	0.62	0.57	0.84	0.59	1.37	0.43	0.78
	0.42	0.56	0.50	0.72	0.57	1.33	0.40	0.74
	0.38	0.55	0.48	0.66	0.54	1.29	0.37	0.69
	0.34	0.49	0.46	0.64	0.48	1.27	0.36	0.62
	0.33	0.43	0.41	0.60	0.44	1.27	0.32	0.61
	0.32	0.41	0.37	0.52	0.41	1.22	0.29	0.59
	0.30	0.39	0.35	0.49	0.40	1.20	0.27	0.53
Δ	6.85	4.87	6.24	6.14	4.79	12.34	5.55	6.29
	0.54	1.20	0.73	1.09	0.73	1.24	0.25	1.01
	0.15	0.13	0.08	0.22	0.04	0.11	0.02	0.16
	0.08	0.20	0.29	0.22	0.16	0.26	0.11	0.10
	0.08	0.13	0.18	0.15	0.04	0.05	0.07	0.07
	0.07	0.06	0.03	0.17	0.21	0.19	0.13	0.17
	0.17	0.08	0.14	0.13	0.08	0.08	0.10	0.16
	0.12	0.13	0.09	0.09	0.13	0.12	0.16	0.12
	0.24	0.21	0.17	0.14	0.05	0.12	0.31	0.01
	0.13	0.04	0.23	0.13	0.52	0.11	0.34	0.32
	0.45	0.59	0.33	0.34	0.20	0.06	0.23	0.08
	0.24	0.11	0.22	0.10	0.08	0.07	0.04	0.18
	0.08	0.11	0.08	0.07	0.08	0.03	0.15	0.15
	0.04	0.06	0.07	0.12	0.02	0.04	0.03	0.04
	0.04	0.01	0.02	0.06	0.03	0.04	0.02	0.05
	0.04	0.07	0.02	0.02	0.06	0.01	0.02	0.07
	0.00	0.06	0.05	0.04	0.05	0.01	0.04	0.01
	0.01	0.02	0.04	0.08	0.02	0.04	0.03	0.02
	0.02	0.02	0.02	0.03	0.02	0.02	0.02	0.05
	0.03	0.03	0.02	0.04	0.03	0.02	0.01	0.04

Note. λ indicates eigenvalue.

Δ indicates change in eigenvalues with each successive component.

First 20 out of 66 eigenvalues are reported.

Appendix C

In this appendix, eigenvalues from the principal components analysis for the second-stage tests for the eight groups of examinees studied in the1999 administration are presented.

	Low-Ability Examinees				High-Ability Examinees			
	13-year-olds		16-year-olds		13-year-olds		16-year-olds	
	English	French	English	French	English	French	English	French
λ	12.39	9.37	11.63	10.33	6.72	6.71	7.96	8.20
	1.27	1.33	1.53	1.71	1.32	1.88	1.24	1.39
	1.10	1.21	1.37	1.48	0.89	1.24	1.12	1.21
	0.88	1.10	1.09	1.24	0.75	1.09	0.69	1.00
	0.71	0.97	0.99	1.14	0.68	0.99	0.56	0.94
	0.61	0.82	0.82	1.04	0.62	0.95	0.52	0.83
	0.57	0.80	0.69	0.99	0.54	0.85	0.49	0.72
	0.53	0.72	0.63	0.89	0.53	0.79	0.48	0.60
	0.47	0.65	0.60	0.85	0.51	0.77	0.42	0.57
	0.44	0.60	0.56	0.83	0.45	0.72	0.39	0.55
	0.42	0.58	0.53	0.75	0.44	0.67	0.38	0.54
	0.40	0.56	0.48	0.69	0.41	0.64	0.36	0.47
	0.39	0.53	0.46	0.69	0.38	0.58	0.34	0.46
	0.34	0.52	0.46	0.62	0.35	0.56	0.33	0.42
	0.31	0.47	0.43	0.61	0.34	0.53	0.29	0.40
	0.31	0.45	0.41	0.57	0.32	0.47	0.27	0.37
	0.28	0.43	0.38	0.55	0.31	0.44	0.26	0.35
	0.27	0.38	0.36	0.53	0.29	0.42	0.26	0.33
	0.26	0.37	0.35	0.48	0.26	0.40	0.25	0.30
	0.26	0.35	0.34	0.46	0.26	0.37	0.24	0.29
Δ	11.12	8.04	10.10	8.62	5.40	4.83	6.73	6.81
	0.17	0.13	0.16	0.23	0.43	0.64	0.12	0.18
	0.22	0.11	0.28	0.24	0.15	0.15	0.43	0.21
	0.17	0.13	0.09	0.10	0.07	0.11	0.13	0.06
	0.10	0.15	0.18	0.10	0.05	0.04	0.04	0.11
	0.05	0.02	0.13	0.06	0.08	0.09	0.03	0.11
	0.03	0.08	0.06	0.10	0.01	0.06	0.01	0.12
	0.06	0.07	0.03	0.03	0.02	0.03	0.06	0.03
	0.03	0.05	0.03	0.03	0.06	0.05	0.02	0.02
	0.02	0.02	0.03	0.08	0.01	0.05	0.02	0.02
	0.02	0.02	0.05	0.06	0.04	0.02	0.01	0.06
	0.01	0.03	0.02	0.00	0.02	0.06	0.02	0.01
	0.05	0.01	0.00	0.06	0.03	0.02	0.01	0.05
	0.02	0.06	0.03	0.01	0.01	0.03	0.04	0.02
	0.00	0.02	0.02	0.04	0.02	0.05	0.02	0.03
	0.03	0.02	0.03	0.02	0.02	0.03	0.01	0.02
	0.01	0.05	0.02	0.03	0.02	0.02	0.01	0.02
	0.01	0.01	0.02	0.04	0.04	0.02	0.01	0.03
	0.00	0.02	0.01	0.03	0.00	0.03	0.01	0.01
	0.01	0.00	0.03	0.01	0.01	0.01	0.02	0.03

Note. λ indicates eigenvalue.
Δ indicates change in eigenvalues with each successive component.
First 20 out of 66 eigenvalues are reported.

Appendix D

The BILOG-MG command files for SAIP Science 1996 test are presented in this appendix.

Analysis 1

```
>COMMENT
Two stage testing for SAIP (First Stage)
>GLOBAL  DFNAME='96fs.DAT', NPARM=3, SAVE;
>SAVE    SCORE='anal1.SCO', PARM='anal1.par';
>LENGTH  NITEMS=(12);
>INPUT   NTOT=12, SAMPLE=24642, NGROUP=8, NIDCH=7, TYPE=1;
>ITEMS   INUM=(1(1)12), INAME=(A01(1)A12);
>TEST    TNAME=FormA,INUM=(1,2,3,4,5,6,7,8,9,10,11,12);
>GROUP1  GNAME=AgroupB1,LENGTH=12,INUM=(1(1)12);
>GROUP2  GNAME=AgroupC2,LENGTH=12,INUM=(1(1)12);
>GROUP3  GNAME=AgroupB3,LENGTH=12,INUM=(1(1)12);
>GROUP4  GNAME=AgroupC4,LENGTH=12,INUM=(1(1)12);
>GROUP5  GNAME=AgroupB5,LENGTH=12,INUM=(1(1)12);
>GROUP6  GNAME=AgroupC6,LENGTH=12,INUM=(1(1)12);
>GROUP7  GNAME=AgroupB7,LENGTH=12,INUM=(1(1)12);
>GROUP8  GNAME=AgroupC8,LENGTH=12,INUM=(1(1)12);
(7A1,I1,12A1)
>CALIB   FIX, NOFLOAT, CYCLE=35, SPRIOR, NEWTON=2, CRIT=0.001, REF=0;
>SCORE   IDIST=3,METHOD=2, NOPRINT, INFO=2, POP;
```

Analysis 2

```
>COMMENT
```

Continued….

Two stage testing for SAIP (Second Stage)

```
>GLOBAL  DFNAME='96ss.DAT', NPARM=2, SAVE;
>SAVE    SCORE='anal2.SCO', PARM='anal2.par';
>LENGTH  NITEMS=(430);
>INPUT   NTOT=430, SAMPLE=24642,NGROUP=8, NFORM=8, NIDCH=7, TYPE=1;
>ITEMS   INUM=(1(1)430),INAME=(A01(1)A430);
>TEST    TNAME='2-STAGE',INUM=(1(1)430);
>FORM1   LENGTH=66,INUM=(1(1)66);
>FORM2   LENGTH=66,INUM=(1(1)14,67(1)118);
>FORM3   LENGTH=66,INUM=(1(1)14,119(1)170);
>FORM4   LENGTH=66,INUM=(1(1)14,171(1)222);
>FORM5   LENGTH=66,INUM=(1(1)14,223(1)274);
>FORM6   LENGTH=66,INUM=(1(1)14,275(1)326);
>FORM7   LENGTH=66,INUM=(1(1)14,327(1)378);
>FORM8   LENGTH=66,INUM=(1(1)14,379(1)430);
>GROUP1  GNAME=1 LENGTH=66,INUM=(1(1)66);
>GROUP2  GNAME=2 LENGTH=66,INUM=(1(1)14,67(1)118);
>GROUP3  GNAME=3 LENGTH=66,INUM=(1(1)14,119(1)170);
>GROUP4  GNAME=4 LENGTH=66,INUM=(1(1)14,171(1)222);
>GROUP5  GNAME=5 LENGTH=66,INUM=(1(1)14,223(1)274);
>GROUP6  GNAME=6 LENGTH=66,INUM=(1(1)14,275(1)326);
>GROUP7  GNAME=7 LENGTH=66,INUM=(1(1)14,327(1)378);
>GROUP8  GNAME=8 LENGTH=66,INUM=(1(1)14,379(1)430);;
(7A1,I1,T8,I1,66A1)
```

Continued....

```
>CALIB NQPT=20, IDIST=1, FIX,NOFLOAT, CYCLE=35, SPRIOR, NEWTON=2,
CRIT=0.001, REF=0,  ACC=0.0 ;
>QUAD1 POINT= (-0.4034E+01 -0.3610E+01 -0.3185E+01 -0.2760E+01 -0.2336E+01
              -0.1911E+01 -0.1486E+01 -0.1062E+01 -0.6371E+00 -0.2124E+00
               0.2123E+00  0.6370E+00  0.1062E+01  0.1486E+01  0.1911E+01
               0.2336E+01  0.2760E+01  0.3185E+01  0.3610E+01  0.4034E+01),
         WEIGHT= (0.1918E-03  0.9026E-03  0.3499E-02  0.1126E-01  0.3022E-01
               0.6797E-01  0.1271E+00  0.1920E+00  0.2227E+00  0.1860E+00
               0.1059E+00  0.4007E-01  0.1014E-01  0.1769E-02  0.2211E-03
               0.2051E-04  0.8705E-14  0.6000E-16  0.3292E-18  0.0000E+00);
>QUAD2 POINT= (-0.4034E+01 -0.3610E+01 -0.3185E+01 -0.2760E+01 -0.2336E+01
              -0.1911E+01 -0.1486E+01 -0.1062E+01 -0.6371E+00 -0.2124E+00
               0.2123E+00  0.6370E+00  0.1062E+01  0.1486E+01  0.1911E+01
               0.2336E+01  0.2760E+01  0.3185E+01  0.3610E+01  0.4034E+01),
         WEIGHT= (0.1458E-03  0.7138E-03  0.2884E-02  0.9742E-02  0.2762E-01
               0.6556E-01  0.1278E+00  0.1972E+00  0.2285E+00  0.1872E+00
               0.1036E+00  0.3796E-01  0.9327E-02  0.1586E-02  0.1943E-03
               0.1777E-04  0.8469E-14  0.5931E-16  0.3304E-18  0.0000E+00);
>QUAD3 POINT= (-0.4034E+01 -0.3610E+01 -0.3185E+01 -0.2760E+01 -0.2336E+01
           -0.1911E+01 -0.1486E+01 -0.1062E+01 -0.6371E+00 -0.2124E+00
               0.2123E+00  0.6370E+00  0.1062E+01  0.1486E+01  0.1911E+01
               0.2336E+01  0.2760E+01  0.3185E+01  0.3610E+01  0.4034E+01),
         WEIGHT= (0.1394E-03  0.6630E-03  0.2596E-02  0.8471E-02  0.2325E-01
               0.5430E-01  0.1079E+00  0.1774E+00  0.2262E+00  0.2064E+00
```

Continued....

```
                 0.1263E+00 0.5041E-01 0.1328E-01 0.2387E-02 0.3057E-03
                 0.2897E-04 0.7386E-14 0.5151E-16 0.2860E-18 0.0000E+00);
>QUAD4 POINT= (-0.4034E+01 -0.3610E+01 -0.3185E+01 -0.2760E+01 -0.2336E+01
                 -0.1911E+01 -0.1486E+01 -0.1062E+01 -0.6371E+00 -0.2124E+00
                 0.2123E+00 0.6370E+00 0.1062E+01 0.1486E+01 0.1911E+01
                 0.2336E+01 0.2760E+01 0.3185E+01 0.3610E+01 0.4034E+01),
          WEIGHT= (0.9081E-04 0.4656E-03 0.1971E-02 0.6961E-02 0.2066E-01
                 0.5168E-01 0.1079E+00 0.1821E+00 0.2328E+00 0.2086E+00
                 0.1239E+00 0.4793E-01 0.1230E-01 0.2176E-02 0.2767E-03
                 0.2627E-04 0.5609E-14 0.3974E-16 0.2238E-18 0.0000E+00);
>QUAD5 POINT= (-0.4034E+01 -0.3610E+01 -0.3185E+01 -0.2760E+01 -0.2336E+01
                 -0.1911E+01 -0.1486E+01 -0.1062E+01 -0.6371E+00 -0.2124E+00
                 0.2123E+00 0.6370E+00 0.1062E+01 0.1486E+01 0.1911E+01
                 0.2336E+01 0.2760E+01 0.3185E+01 0.3610E+01 0.4034E+01),
 WEIGHT= (0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.8875E-04
                 0.6711E-03 0.4216E-02 0.2020E-01 0.6789E-01 0.1512E+00
                 0.2208E+00 0.2193E+00 0.1588E+00 0.9014E-01 0.4227E-01
                 0.1675E-01 0.5631E-02 0.1598E-02 0.3798E-03 0.7336E-04);
>QUAD6 POINT= (-0.4034E+01 -0.3610E+01 -0.3185E+01 -0.2760E+01 -0.2336E+01
                 -0.1911E+01 -0.1486E+01 -0.1062E+01 -0.6371E+00 -0.2124E+00
                 0.2123E+00 0.6370E+00 0.1062E+01 0.1486E+01 0.1911E+01
                 0.2336E+01 0.2760E+01 0.3185E+01 0.3610E+01 0.4034E+01),
          WEIGHT= (0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1346E-03
                 0.9446E-03 0.5504E-02 0.2453E-01 0.7715E-01 0.1621E+00
```

Continued....

```
          0.2255E+00  0.2151E+00  0.1502E+00  0.8205E-01  0.3684E-01
          0.1394E-01  0.4485E-02  0.1225E-02  0.2818E-03  0.5272E-04);
>QUAD7 POINT= (-0.4034E+01 -0.3610E+01 -0.3185E+01 -0.2760E+01 -0.2336E+01
          -0.1911E+01 -0.1486E+01 -0.1062E+01 -0.6371E+00 -0.2124E+00
          0.2123E+00  0.6370E+00  0.1062E+01  0.1486E+01  0.1911E+01
          0.2336E+01  0.2760E+01  0.3185E+01  0.3610E+01  0.4034E+01),
      WEIGHT=  (0.0000E+00  0.0000E+00  0.0000E+00  0.0000E+00  0.5729E-04
          0.4429E-03  0.2888E-02  0.1456E-01  0.5213E-01  0.1254E+00
          0.2006E+00  0.2212E+00  0.1785E+00  0.1120E+00  0.5690E-01
          0.2387E-01  0.8331E-02  0.2424E-02  0.5869E-03  0.1153E-03);
>QUAD8 POINT= (-0.4034E+01 -0.3610E+01 -0.3185E+01 -0.2760E+01 -0.2336E+01
          -0.1911E+01 -0.1486E+01 -0.1062E+01 -0.6371E+00 -0.2124E+00
          0.2123E+00  0.6370E+00  0.1062E+01  0.1486E+01  0.1911E+01
          0.2336E+01  0.2760E+01  0.3185E+01  0.3610E+01  0.4034E+01),
      WEIGHT=  (0.0000E+00  0.0000E+00  0.0000E+00  0.0000E+00  0.8637E-04
          0.6209E-03  0.3770E-02  0.1779E-01  0.6007E-01  0.1373E+00
          0.2099E+00  0.2214E+00  0.1703E+00  0.1014E+00  0.4886E-01
          0.1952E-01  0.6548E-02  0.1848E-02  0.4373E-03  0.8422E-04);
      >SCORE   IDIST=3,METHOD=2, NOPRINT, INFO=2, POP;
```

Appendix E

The BILOG-MG command files for SAIP Science 1996 test are presented in this appendix.

Analysis 1

```
>COMMENT
Two-stage testing for SAIP (first-stage analysis)
>GLOBAL  DFNAME='merged.DAT', NPARM=3, SAVE;
>SAVE    SCORE='anal1.SCO', PARM='anal1.par';
>LENGTH  NITEMS=(12);
>INPUT   NTOT=12, SAMPLE=22330, NGROUP=8, NIDCH=7, TYPE=1;
>ITEMS   INUM=(1(1)12), INAME=(A01(1)A12);
>TEST    TNAME=FormA,INUM=(1,2,3,4,5,6,7,8,9,10,11,12);
>GROUP1  GNAME=AgroupB1,LENGTH=12,INUM=(1(1)12);
>GROUP2  GNAME=AgroupC2,LENGTH=12,INUM=(1(1)12);
>GROUP3  GNAME=AgroupB3,LENGTH=12,INUM=(1(1)12);
>GROUP4  GNAME=AgroupC4,LENGTH=12,INUM=(1(1)12);
>GROUP5  GNAME=AgroupB5,LENGTH=12,INUM=(1(1)12);
>GROUP6  GNAME=AgroupC6,LENGTH=12,INUM=(1(1)12);
>GROUP7  GNAME=AgroupB7,LENGTH=12,INUM=(1(1)12);
>GROUP8  GNAME=AgroupC8,LENGTH=12,INUM=(1(1)12);
(7A1,I1,12A1)
>CALIB   FIX, NOFLOAT, CYCLE=35, SPRIOR, NEWTON=2, CRIT=0.001, REF=0;
>SCORE   IDIST=3,METHOD=2, NOPRINT, INFO=4, POP;
```

Analysis 2

```
>COMMENT
```

Continued....

Two stage testing for SAIP (second-stage analysis)

```
>GLOBAL  DFNAME='new.DAT', NPARM=2, SAVE;
>SAVE    SCORE='anal2.SCO', PARM='anal2.par';
>LENGTH  NITEMS=(430);
>INPUT   NTOT=430, SAMPLE=22330,NGROUP=8, NFORM=8, NIDCH=7, TYPE=1;
>ITEMS   INUM=(1(1)430),INAME=(A01(1)A430);
>TEST    TNAME='2-STAGE',INUM=(1(1)430);
>FORM1   LENGTH=66,INUM=(1(1)66);
>FORM2   LENGTH=66,INUM=(1(1)14,67(1)118);
>FORM3   LENGTH=66,INUM=(1(1)14,119(1)170);
>FORM4   LENGTH=66,INUM=(1(1)14,171(1)222);
>FORM5   LENGTH=66,INUM=(1(1)14,223(1)274);
>FORM6   LENGTH=66,INUM=(1(1)14,275(1)326);
>FORM7   LENGTH=66,INUM=(1(1)14,327(1)378);
>FORM8   LENGTH=66,INUM=(1(1)14,379(1)430);
>GROUP1  GNAME=1 LENGTH=66,INUM=(1(1)66);
>GROUP2  GNAME=2 LENGTH=66,INUM=(1(1)14,67(1)118);
>GROUP3  GNAME=3 LENGTH=66,INUM=(1(1)14,119(1)170);
>GROUP4  GNAME=4 LENGTH=66,INUM=(1(1)14,171(1)222);
>GROUP5  GNAME=5 LENGTH=66,INUM=(1(1)14,223(1)274);
>GROUP6  GNAME=6 LENGTH=66,INUM=(1(1)14,275(1)326);
>GROUP7  GNAME=7 LENGTH=66,INUM=(1(1)14,327(1)378);
>GROUP8  GNAME=8 LENGTH=66,INUM=(1(1)14,379(1)430);;
(7A1,I1,T8,I1,66A1)
```

Continued….

```
>CALIB NQPT=20, IDIST=1, FIX,NOFLOAT, CYCLE=35, SPRIOR, NEWTON=2,
CRIT=0.001, REF=0,  ACC=0.0 ;
>QUAD1 POINT= (-0.4034E+01 -0.3610E+01 -0.3185E+01 -0.2760E+01 -0.2336E+01
        -0.1911E+01 -0.1486E+01 -0.1062E+01 -0.6371E+00 -0.2124E+00
         0.2123E+00  0.6370E+00  0.1062E+01  0.1486E+01  0.1911E+01
         0.2336E+01  0.2760E+01  0.3185E+01  0.3610E+01  0.4034E+01),
    WEIGHT= (0.2253E-03  0.1069E-02  0.4149E-02  0.1329E-01  0.3525E-01
         0.7780E-01  0.1416E+00  0.2057E+00  0.2246E+00  0.1720E+00
         0.8756E-01  0.2913E-01  0.6449E-02  0.9881E-03  0.1098E-03
         0.8571E-05  0.2298E-14  0.1383E-16  0.6655E-19  0.0000E+00);
>QUAD2 POINT= (-0.4034E+01 -0.3610E+01 -0.3185E+01 -0.2760E+01 -0.2336E+01
        -0.1911E+01 -0.1486E+01 -0.1062E+01 -0.6371E+00 -0.2124E+00
         0.2123E+00  0.6370E+00  0.1062E+01  0.1486E+01  0.1911E+01
         0.2336E+01  0.2760E+01  0.3185E+01  0.3610E+01  0.4034E+01),
    WEIGHT=  (0.1452E-03  0.7331E-03  0.3042E-02  0.1055E-01  0.3063E-01
         0.7359E-01  0.1420E+00  0.2112E+00  0.2302E+00  0.1741E+00
0.8753E-01  0.2880E-01  0.6301E-02  0.9521E-03  0.1040E-03
         0.7937E-05  0.1218E-14  0.7323E-17  0.3520E-19  0.0000E+00);
>QUAD3 POINT= (-0.4034E+01 -0.3610E+01 -0.3185E+01 -0.2760E+01 -0.2336E+01
        -0.1911E+01 -0.1486E+01 -0.1062E+01 -0.6371E+00 -0.2124E+00
         0.2123E+00  0.6370E+00  0.1062E+01  0.1486E+01  0.1911E+01
         0.2336E+01  0.2760E+01  0.3185E+01  0.3610E+01  0.4034E+01),
    WEIGHT= (0.1360E-03  0.6480E-03  0.2535E-02  0.8273E-02  0.2284E-01
         0.5436E-01  0.1111E+00  0.1864E+00  0.2357E+00  0.2060E+00
```

Continued....

```
      0.1172E+00 0.4271E-01 0.1017E-01 0.1653E-02 0.1926E-03
      0.1599E-04 0.1014E-14 0.5950E-17 0.2795E-19 0.0000E+00);
>QUAD4 POINT= (-0.4034E+01 -0.3610E+01 -0.3185E+01 -0.2760E+01 -0.2336E+01
      -0.1911E+01 -0.1486E+01 -0.1062E+01 -0.6371E+00 -0.2124E+00
      0.2123E+00 0.6370E+00 0.1062E+01 0.1486E+01 0.1911E+01
      0.2336E+01 0.2760E+01 0.3185E+01 0.3610E+01 0.4034E+01),
   WEIGHT= (0.8108E-04 0.4253E-03 0.1822E-02 0.6549E-02 0.1988E-01
      0.5131E-01 0.1104E+00 0.1890E+00 0.2395E+00 0.2084E+00
      0.1180E+00 0.4273E-01 0.1010E-01 0.1627E-02 0.1876E-03
      0.1535E-04 0.5226E-15 0.3087E-17 0.1457E-19 0.0000E+00);
>QUAD5 POINT= (-0.4034E+01 -0.3610E+01 -0.3185E+01 -0.2760E+01 -0.2336E+01
      -0.1911E+01 -0.1486E+01 -0.1062E+01 -0.6371E+00 -0.2124E+00
      0.2123E+00 0.6370E+00 0.1062E+01 0.1486E+01 0.1911E+01
      0.2336E+01 0.2760E+01 0.3185E+01 0.3610E+01 0.4034E+01),
 WEIGHT= (0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.1288E-03
      0.9650E-03 0.5832E-02 0.2619E-01 0.8126E-01 0.1667E+00
      0.2260E+00 0.2112E+00 0.1459E+00 0.7976E-01 0.3614E-01
      0.1384E-01 0.4505E-02 0.1226E-02 0.2886E-03 0.5476E-04);
>QUAD6 POINT= (-0.4034E+01 -0.3610E+01 -0.3185E+01 -0.2760E+01 -0.2336E+01
      -0.1911E+01 -0.1486E+01 -0.1062E+01 -0.6371E+00 -0.2124E+00
      0.2123E+00 0.6370E+00 0.1062E+01 0.1486E+01 0.1911E+01
      0.2336E+01 0.2760E+01 0.3185E+01 0.3610E+01 0.4034E+01),
   WEIGHT= (0.0000E+00 0.0000E+00 0.0000E+00 0.0000E+00 0.2158E-03
      0.1456E-02 0.7970E-02 0.3280E-01 0.9467E-01 0.1829E+00
```

Continued....

```
          0.2343E+00  0.2057E+00  0.1320E+00  0.6640E-01  0.2769E-01
          0.9869E-02  0.3036E-02  0.7871E-03  0.1811E-03  0.3348E-04);
>QUAD7 POINT= (-0.4034E+01 -0.3610E+01 -0.3185E+01 -0.2760E+01 -0.2336E+01
          -0.1911E+01 -0.1486E+01 -0.1062E+01 -0.6371E+00 -0.2124E+00
          0.2123E+00  0.6370E+00  0.1062E+01  0.1486E+01  0.1911E+01
          0.2336E+01  0.2760E+01  0.3185E+01  0.3610E+01  0.4034E+01),
    WEIGHT= (0.0000E+00  0.0000E+00  0.0000E+00  0.0000E+00  0.6961E-04
          0.5525E-03  0.3569E-02  0.1729E-01  0.5859E-01  0.1332E+00
          0.2033E+00  0.2169E+00  0.1719E+00  0.1070E+00  0.5417E-01
          0.2268E-01  0.7903E-02  0.2281E-02  0.5571E-03  0.1105E-03);
>QUAD8 POINT= (-0.4034E+01 -0.3610E+01 -0.3185E+01 -0.2760E+01 -0.2336E+01
          -0.1911E+01 -0.1486E+01 -0.1062E+01 -0.6371E+00 -0.2124E+00
0.2123E+00  0.6370E+00  0.1062E+01  0.1486E+01  0.1911E+01
          0.2336E+01  0.2760E+01  0.3185E+01  0.3610E+01  0.4034E+01),
    WEIGHT=  (0.0000E+00  0.0000E+00  0.0000E+00  0.0000E+00  0.1054E-03
          0.7685E-03  0.4562E-02  0.2044E-01  0.6475E-01  0.1397E+00
          0.2054E+00  0.2137E+00  0.1665E+00  0.1021E+00  0.5098E-01
          0.2107E-01  0.7261E-02  0.2079E-02  0.5045E-03  0.9950E-04);
>SCORE   IDIST=3,METHOD=2, NOPRINT, INFO=4, POP;
```